IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 July 2024, accepted 6 August 2024, date of publication 22 August 2024, date of current version 3 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3448247

RESEARCH ARTICLE

Navigating the Shadows: Manual and Semi-Automated Evaluation of the Dark Web for Cyber Threat Intelligence

PHILIPP KÜHN, KYRA WITTORF, AND CHRISTIAN REUTER
Science and Technology for Peace and Security (PEASEC), Technical University of Darmstadt, 64289 Darmstadt, Germany

Corresponding author: Philipp Kühn (kuehn@peasec.tu-darmstadt.de)

This work was supported in part by German Federal Ministry of Education and Research (BMBF) in the Project CYWARN under Grant 13N15407; and in part by German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts in their Joint Support of the ATHENE National Research Center for Applied Cybersecurity.

ABSTRACT In today's world, cyber-attacks are becoming more frequent and thus proactive protection against them is becoming more important. Cyber Threat Intelligence (CTI) is a possible solution, as it collects threat information in various information sources and derives stakeholder intelligence to protect one's infrastructure. The current focus of CTI in research is the clear web, but the dark web may contain further information. To further advance protection, this work analyzes the dark web as Open Source Intelligence (OSINT) data source to complement current CTI information. The underlying assumption is that hackers use the dark web to exchange, develop, and share information and assets. This work aims to understand the structure of the dark web and identify the amount of its openly available CTI related information. We conducted a comprehensive literature review for dark web research and CTI. To follow this up we manually investigated and analyzed 65 dark web forum (DWF), 7 single-vendor shops, and 72 dark web marketplace (DWM). We documented the content and relevance of DWFs and DWMs for CTI, as well as challenges during the extraction and provide mitigations. During our investigation we identified IT security relevant information in both DWFs and DWMs, ranging from *malware toolboxes* to *hacking-as-a-service*. One of the most present challenges during our manual analysis were necessary interactions to access information and anti-crawling measures, *i.e.*, CAPTCHAs. This analysis showed 88% of marketplaces and 53% of forums contained relevant data. Our complementary semi-automated analysis of 1 186 906 onion addresses indicates, that the necessary interaction makes it difficult to see the dark web as an open, but rather treat it as specialized information source, when clear web information does not suffice.

INDEX TERMS Dark web, investigation, security.

I. INTRODUCTION

Information technology becomes more ubiquitous with every day, and with it, the risks of cyberattacks and cybercrime grows. In 2023, the estimated costs of global cybercrime have reached 8.15 trillion dollars [2], driven by hackers who continuously develop novel techniques to attack IT systems and personnel [1], [3], [4], [5], [6], [7], [8]. Those attacking techniques are shared on the web by attackers in their communities. This includes vulnerabilities, hacking tools, malware, and hacking services [4]. The dark web, in particular, plays a crucial role in this process.

The web comprises three layers: the surface web, accessible via ordinary browsers and search engines, the deep web, which includes non-indexed pages that are usually only accessible through authentication processes, *e.g.*, internal networks, and the dark web, which can provides encryption and anonymity through protocols [9]. Both, surface- and deep web provide no anonymity, making these parts of the web less attractive to criminal hackers [10], [11], [12], [13]. Only their availability is still counting towards those services [13]. The dark web, which provides criminals a space for their

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleylek.

activities provided in Hidden Service (HS), is a crucial source of intelligence for understanding and mitigating cyber threats [3], [12], [14].

Hence, it is necessary to establish new protection measures and derive information about potential attacks, which is one of the key goals of Cyber Threat Intelligence (CTI). The idea in this work is to provide an overview and analysis of current information sources in the dark web, whether they can be used to obtain such information, as the benefits of open information sources cannot be denied [15]. Hence, potentially relevant information can be found on the dark web. In addition, the dark web may be an early source of information, as attacks are planned here so that countermeasures can be developed as early as possible based on this information.

A. GOAL

The focus in this work is on gaining knowledge of the benefits of the dark web for IT security. In doing so, the The Onion Router (Tor) network is considered as it is basically a gateway to the dark web. The Tor network is the best known and largest network.

During our investigation, we plan to answer the following research questions: *"How can IT security-related information, such as vulnerabilities and exploits, be identified on the dark web? **(RQ1)**"*, *"What challenges exist in extracting and analyzing open dark web data? **(RQ2)**"*. *"To what extent can the dark web be utilized for CTI? **(RQ3)**"*, and *"What is the nature and volume of IT security information on the dark web, and how does it contribute to CTI landscape? **(RQ4)**"*.

This work makes several contributions to CTI and the dark web research. First, it provides a *manual analysis of forums and marketplaces* provides a new perspective for automated CTI research and is complemented by a complementary *semi-automated analysis of HS* **(C1)**. The other contribution *outlines current challenges for automatically gathering open information on the dark web*, with a focus on overcoming CAPTCHAs **(C2)**.

The paper is structured as follows: §II reviews related work, followed by the methodology (*cf.* §III). Our analysis is presented in §IV. We discuss the results in §V and conclude our paper in §VI with future work directions.

II. RELATED WORK

This section provides an overview of previous work on crawling and research of the dark web in general, followed by CTI and vulnerability databases in combination with the dark web. The review process is conducted using commonly known digital libraries[1] and a snowball sampling process afterward. We conclude this section with a research gap.

Automated extraction of information from the web relies on web crawlers. This is no different for the dark web, except the underlying protocol to connect to this part of the web. [16] presents an overview of current research in dark web crawling. It outlines several currently used methods to mitigate hurdles, *e.g.*, semi-automation via human interactions. We outline additional information of current research in Table 1. The present work on the other hand aims for a qualitative approach based on a manual analysis combined with a semi-automated analysis, rather than (semi-)automated crawling approach and computational classification.

Others delve into a variety of dark web themes, including drug trafficking, hacker activities, and terrorism. The Darknet Identification, Collection, Evaluation with Ethics (DICE-E) framework [17] emphasizes ethical considerations, guiding researchers in safe and ethical dark web exploration. [18] presents problems and further mitigations and recommendations. All those studies not only contribute to academic knowledge but also aid in developing strategies to counter cybercrime and terrorism effectively. Compared to them, we focus mainly on CTI information. Reference [19] present another survey of current dark web research.

Researchers explore various aspects of the dark web, including asset exchange (tools, exploits) for cyberattacks [20]. Research methods encompass supervised Machine Learning (ML) for cyber-security relevance classification of text, with Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) utilized [21], [22], [23], [24], [25]. Multinomial classification categorizes forum posts into nine classes for cyber threat extraction [23]. Unsupervised ML, such as clustering, identifies topics, and product categories [26], [27]. Researchers also detect new threats using terms associated with malware, ransomware, data leaks, or vulnerabilities [4], [28]. Additionally, predictive models aim to forecast cyberattacks, treated as binary classification problems [29], [30]. Social Network Analyses (SNAs) offer deep insights into dark web communities, enabling the analysis of social connections and actor identification [12], [27], [31], [32]. Source code analysis classifies attachments and tutorials for programming languages and exploit types [20], [33], [34], [35]. Studies vary in data collection methods, some utilizing existing data [21], [22], [26], [36], [37], while others involve active crawling [12], [20], [21], [25], [28], [33], [38], [39], [40]. The research scope ranges from forums to marketplaces, with some covering multiple platforms concurrently [4], [12], [17], [22], [35], [41], [42], [43]. These results are still in line with the results identified by [44]. In comparison to related work, our approach is rather holistic in nature. Our objective is to identify all information that is somehow related to cybersecurity and accompanying threats.

Current research on how dark web information can be used to enhance information in vulnerability databases, remains limited [21], [29], [30], [45]. One objective of this research is to predict cyber threats, often framed as binary classification problems, considering whether vulnerabilities are exploited [29], [30], [40], [45], [46]. Common Vulnerabilities and Exposure (CVE) data, crucial for cybersecurity,

[1] IEEExplore, ACM Digital Library, arXiv, Google Scholar, and Semantic Research.

serves as a key parameter in this analysis [4], [21], [23], [47]. Some studies employ explicit CVE mentions in posts, using techniques from SNA and ML like RF to correlate hacking activities with vulnerabilities and real-world cyber incidents [30], [46]. Others expand the scope by considering posts mentioning vulnerabilities or extracting Common Platform Enumeration (CPE) data for a multi-label classification approach, utilizing models like SVM and RF [21]. The [1] introduces an integrated framework that monitors data sources, including National Vulnerability Database (NVD) and Twitter, considering CVE as a critical parameter. It aids in identifying, collecting, analyzing, extracting, integrating, and sharing CTI [1]. Additionally, research efforts strive to link exploits to vulnerabilities by extracting exploit and vulnerability names, *e.g.*, a vulnerability identifier (CVE-2014-0160) or their given name (here *Heartbleed*), from dark web forums, utilizing advanced techniques such as Bi-Long Short Term Memory (LSTM) and attention mechanisms [48]. Moreover, a novel vulnerability severity metric is developed, combining vulnerability information with data from linked hacker exploits [48]. As stated, we aim to present a holistic view of available CTI information for the selected marketplaces and forums, based on a manual, rather than automatic analysis.

A. RESEARCH GAP

The benefits of Open Source Intelligence (OSINT) sources for information gathering can not be denied, and the dark web needs to be analyzed further in this regard [15]. Research on the dark web as a potentially relevant source for CTI is researched in various papers. However, current research looks for specific information (tool, vulnerabilities, exploits, or posts) rather than trying to provide a thorough overview of information, which is one of the key aspects of CTI [1], [20], [21], [63]. For data gathering, researchers usually use page-specific crawlers with manual interventions, *e.g.*, manually solving Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) and provide the session tokens to the crawler [61]. Others [36] use pre-compiled datasets to assess the availability if CTI information, without assessing the gathering process as a whole. A thorough analysis of the suitability of the dark web for current CTI information gathering using solely open information and assessing the problems for (partially) automated crawlers and possible mitigations is not done as far as we know.

III. METHODOLOGY

This section describes the used methodology to conduct our research. First we specify the methodology for the manual investigations part, followed by the semi-automatic analysis. In addition, we stated the ethical considerations during our research to protect ourselves from harmful content. The whole research process is conducted by two information security researches. In total, we manually analyze 65 dark web forums (DWFs), 7 single-vendor shops, and 72 dark web marketplaces (DWMs). The semi-automatic analysis is based on a provided database snapshot timestamped at 26th June 2022.

A. INVESTIGATION

In §II a comprehensive literature review has already been conducted, which forms the information basis. Based on these papers, various marketplaces and forums are examined manually. Hereby, we aim for a qualitative analysis, rather than a quantitative [69]. Implementing and maintaining handwritten crawlers is possible during the course of a study, they quickly become out of date afterward, due to the dynamics of the dark web, hindering the reproducibility of the whole study. Moreover, while automated approaches seem to dominate, it is not an uncommon way to approach this field [69]. In order to gain as much insight as possible into the dark web for IT security, it is necessary to manually visit as many HS as possible. For this, preventive security measures are necessary to ensure the safe browsing experience and download of the web pages. Further, the scan is run in a virtualized environment so that an infection cannot spread to other resources. At the same time, JavaScript is turned off by default in the browser so that unwanted JavaScript code is not executed [17]. To get potentially relevant HS entry points into the dark web, we use various clearnet websites[2] and filter them for security related links. Those pages tag links based on their provided content. As the work by [70] suggest, we filter for *digital* as well as *security*.. Following this, we contacted these entry points on three separate occasions: November 2022, shortly afterward, and February 2023, aiming to detect possible downtimes. In the fully manual investigation, the search is limited to DWF, single-vendor shops, and DWM, which are accessed after collection. In addition, the Distributed Denial-of-Service (DDoS) mechanism is documented, and user accounts are created if necessary. This is often a prerequisite for exploring the site, especially for DWFs and DWMs. A user account rarely requires a valid e-mail address on the dark web. Such email addresses are created in advance for the registration process. During the inspection process for the login and registration process, the protection mechanism and the requirements for entering the site, such as an invitation or payment, are documented. For all emerging CAPTCHAs a screenshot is taken to understand the limitation and variations of the mechanisms and to support future research in this area. To access additional information from the page, the operator may require further actions, such as solving programming challenges or achieving a higher level of trust through interactions. The documented information should provide insight into the extent to which the information is publicly available and the hurdles that must be overcome for automatic extraction. For information extraction, we navigate through the (page) menu and document, which categories and

[2] https://torhoo.com/, https://tor.fish/, https://darktrain.express/, and https://darkurl.site/

TABLE 1. Related work on crawling: Empty fields are not clearly outlined in the paper. Source states if either the surface web is used (○), the dark web is used (●), or both (◐). Focus states if it is a focused (✓) or unfocused (✗). The frequency (Freq.) is either one-time (·), periodically (∞), or incremental (↗). The queue states whether a breath-first search queue (↔), depth-first search queue (↓), or a combination of both (⇹) is used. The architecture is given as either centralized (⊙) or parallel (∥). CAPTCHAs are either handled manually (⊥), automatically (⊤), or are ignored or not necessary to deal with (✗).

	Goal	Source	Focus	Freq.	Queue	Arch	CAPTCHA	Tools
[25]	IoT-related CTI	◐	✓	·	↓	⊙	⊥	ACHE, elasticsearch
[49]	Extremist forums	●	✓	↗	⇹	∥	⊥	
[32]	Use free tools	●	✗	·		⊙	✗	AppleScript
[50]	Categorize HS	●	✗	·		∥	✗	
[38]	Databases	○	✓	↗		∥	⊥	Selenium, Bash
[21]	Hacking info	◐	✓					
[47]	CTI	●	✓					
[51]	Products	●	✗	∞		∥	⊤	Python
[52]	Adaptive crawler	◐	✓		↔	⊙		Apache Nutch
[43]	CTI	●		↗				
[53]	State of HS	●	✗	∞		∥		Docker, Selenium, Cloud
[33]	CTI	●	✗	↗	↓	⊙		BeautifulSoup, keras
[54]	Various	●					⊥	
[1]	IoT-based CTI	◐	✓	↗	↔	∥		Selenium, ACHE
[55]	Marketplaces	●						Scrapy, Privoxy
[56]	Agora	●	✗	↗			⊥	Docker, PHP, cURL
[57]	Drugs	●	✗	·	⇹		⊥	Selenium, AWS
[58]	HS	●	✗					Scrapy, Selenium
[59]	Anomaly	●			↔		⊥	
[31]	Criminal & illicit content	●	✓			∥		Privoxy
[60]	Captcha Breaking	●				∥	⊤	Selenium
[61]	Drugs	●					⊥	
[62]	Measurement of DWM size	●	✗	∞	↓	∥	⊥	

subcategories are security relevant and how many entries they contain. Hereby, we follow the keyword list in Fig. 1. This list also includes keywords like *phishing* which are part of two worlds, namely, CTI, as it might focus one's organization, and cybercrime. Such an overlap might occur in the other areas of the present work as well. After that, various CTI-relevant categories are browsed, in order to gain impressions of IT security events. For specific information, *e.g.*, to locate CVE mentions, we use website searches when available. Other information that we look out for are: offerings of hacking services, scamming, or social engineering (incl. phishing). Likewise, FAQ and profile settings are visited to gain more information if necessary. In DWF, the first page of threads is visited in relevant categories and the first page of posts is visited for relevant threads. When browsing we document the IT security relevance, occurrences of specific topics, special features, and CVE mentions. The relevance is documented as either *high*, *medium*, or *low*. In DWM, we visit the first page of each relevant category. We browse relevant products, categories, CVE mentions, and document their IT security relevance. To achieve traceability of the research, we store all visited HTML pages. Additionally, the assessment of the IT security relevance is made by two researchers, proficient in the field of IT security, and differences during the assessment are discussed in case of a tie. In doing so, a domain is assigned to one of the three relevance classes *high*, *medium*, or *low* based on the number, variety, and strength of the relevant information. The assessment includes parameters like (i) availability of the service, (ii) number of posts or products related to IT security, (iii) number of posts or products overall, or (iv) policies regarding hacking or IT security. If a domain contains no signification amount of information related to IT security ($\leq$ 10 or $\leq$ 0.1 % posts/products), it is classified as *low*. If it contains only a few posts or products ($>$ 10 or $>$ 0.1 %) and allows hacking related communication, it is classified as *middle*. And if a page contains many posts, either in absolute or relative amount ($>$ 500 or $>$ 30 %) it is classified as *high*. It should be noted that not all products/posts are reviewed, but only the first page of each (sub)category, thread, or board. A full analysis is considered part of automated procedures and automated processing due to the amount of data involved and leads to different problems, *e.g.*, automatic CAPTCHA-breaking, dealing with automatic logins, generating interactions, and solving other challenges.

> 0-Day, Zero-Day, Zeroday, 0day, Exploit, Malware, rootkit, snort, yara, clamav, DDoS, DoS, Denial-of-Service, denial of service, out-of-service, Ransomware, ransom, Hacking, hacked, Vulnerability, vulnerabilities, vuln, vulns, Botnet, bots, c2 bots, c&c, command and control, Attack, SQL Injection, Keylogger, Phishing, trojan, virus, XSS, cross-scripting, compromise

FIGURE 1. IT security keyword list.

TABLE 2. Overview of CTI from the dark web based on the goal of the research, used tools, and source of the dataset. The sources is a combination of boards (B), websites (P), marketplaces (M), chats (C), or social media (S). Unfilled fields are not shown in the corresponding paper.

	Goal	Tools	Source
[25]	CTI relevance ranking	Gensim, SVM	BMP
[22]	Classify hacking related info	Supervised learning	M
[64]	Classify data breach	Supervised learning	B
[21]	Analysis tool of DWF	Supervised learning	BM
[12]	Board analysis		B
[26]	Hacker DWF analysis	SVM, LDA	B
[27]	Comprehend dealer behavior on DWM	K-Means, SNA	M
[65]	CTI with exploit focus		B
[66]	DWM and DWF information link	SNA, gephi	BM
[40]	Current threat campaigns	SVM, SGNS, RF	
[67]	Correlate threats and dark web discussions	RegEx, LP, PFR	BM
[37]	Threat prediction	LSTM, NER, RF	B
[42]	CTI	CL-LSTM	M
[28]	CTI	Elasticsearch	BMS
[41]	Key players, social connections	SNA	B
[34]	Hacker tools, their key players and source code, CTI	SVM, LDA, SNA	B
[17]	Analyze hacker community		BCM
[23]	Detect CTI in DWF	CNN, SVM, DT, BOW, n-gram	B
[35]	Analyze malware	LSTM, RNN	B
[43]	Integrate various sources in CTI platform	RegEx	BCM
[33]	Visualize cyber threats	RNN, LSTM	B
[1]	Collect and analyze IoT CTI	CNN, NLP, ML, NER	BMP
[20]	Collect and analyze exploit source code	DTL-EL	BM
[48]	Collect exploit source code and link to CVE	LSTM, DSSM	B
[4]	Predict cyber threats	MLP	M
[29]	Predict cyber threats	SNA, ML, RF	B
[30]	Predict incoming cyber threats for organizations	Supervised learning	B
[46]	Predict exploit usage	Supervised learning	BM
[68]	Early warnings of cyber threats	RegEx	B
[45]	Correlate cyber activities and current threats		B
[59]	Automated threat detection	LSTM	M

B. SEMI-AUTOMATIC ANALYSIS

In order to achieve a higher diversity of HS and to integrate different types of HS in this work, a further step is to visit more HS in a semi-automated way. The goal of this step is to get an impression about the IT security related content and its page types of the dark web. The onion addresses are extracted from an existing dark web database generated by the *Panda Projekt*[3] using a keyword list. This database is generated by regular automatic crawls of the Tor network [71], [72], [73]. Fig. 5 visualizes this process.

The non-all-inclusive keyword list includes various IT security-related keywords including their spelling mistake counterparts *cf.* Fig. 1. We extracted 1 186 906 onion addresses this way. During preprocessing, version 2 onion addresses (≈ 563 000) and duplicates (≈ 298 000) are removed, resulting in 5 560 unique base domains containing all remaining addresses, *i.e.*, 5 560 distinct pages to continue analyzing. For the time being, we restrict ourselves to the total of 5 560 base domains for further manual analysis and do not consider subdomains and paths, because of the high number of paths and subdomains and the associated complexity. This seems to be a good abstraction, since on the one hand subdomains should often match the category of the main domain and on the other hand the paths sometimes refer to JavaScript files that do not reflect the content of the page in a meaningful way. After preprocessing the onion addresses, all remaining addresses are visited automatically to get an overview of their raw and textual content, which is then stored in the database. During automatic crawling, Privoxy[4] is used to connect to the Tor network. CAPTCHAs are not handled during this process.

The onion addresses are not tracked in depth and thus only one level of breadth-first search is performed. We save the raw HTML content and the extracted textual content. By saving the textual content in this way, it is possible to obtain a lot of information about the website without having to visit it manually. In addition, this protects the researchers from visiting pornographic web pages, since they get a textual overview beforehand.[5] Such web pages are removed at this examination level based on a keyword list.[6] All HS that are inaccessible during this crawl are crawled again after some time, to catch, whether it was only a short outage. Afterward, the content of the page is analyzed. At the same time, duplicates with the same content, but different domain, are removed. All other textual content is manually examined

[3] https://panda-projekt.de/

[4] https://www.privoxy.org/

[5] Other non-CTI-related content, such as drugs or weapons, were not considered harmful by and for the authors.

[6] We compiled a pornographic keyword list based on https://relatedwords.io/pornographic and added additional words which this list missed: porn, masturbat, lesbian, penetration, pussy, rape, sex, erotic, anal, pedo, incest, dick, gangbang, cock, foreskin, naked, glan, vagina.

and, if it indicates promising results, the website is visited. The relevance assessment (*high*, *middle*, *low*) follows the same procedure as in the manual analysis.

C. ETHICS

We explicitly provide the ethical viewpoint for the present work based on work by [74], [75], and [76]. Reference [77] present further information on critical aspect of darknet research.[7] We provide an overview of why we consider our work research on open information. All HS investigated are publicly accessible on the dark web. If necessary, accounts are created using cost-free dark web email providers, accessible to all. These accounts access public information without interacting with the community or operators, ensuring no disruption or hacking occurs. No posts, purchases, or payments are made, and server operations remain unaffected. This principle follows previous research standards [60], [69], [78]. Reference [78] states: "The data we collected is essentially public. We did have to create an account on Silk Road to access it; but registration is open to anybody who connects to the site. We did not compromise the site in any way". Furthermore, only data that does not contain personal information and is also publicly accessible, such as product information or discussions in forums, is examined in this work. We do not make the data publicly available, as it may contain personally identifiable information or metadata, which, combined with other data, could result in re-identification of individual.

IV. ANALYSIS

We present the results of manual analysis of DWFs and DWMs and the semi-automated extraction of information from various onion domains. First we outline the used sources and their benefits for IT security *cf.* §IV-A. These were collected as part of the literature review and manual exploration process. We follow up with IT security relevant information in these sources *cf.* §IV-B. Additionally, we discuss the relevance of HS. We conclude this section by highlighting challenges of Information Retrieval (IR) in the dark web *cf.* §IV-C. The listed challenges are either part of the manual analysis or the automated crawl.

A. INFORMATION SOURCES

The Tor network provides so-called HS, which hide the IP address of the server. There are different types of HS, *e.g.*, DWM, single-vendor stores, DWF, and chats. Not all HS are illegal or have added value to IT security. From [79] it appears that out of 1 000 HS, approximately 68 % had illegal content.

- **Dark web marketplaces** facilitate product and service trade [21], [66], with specialized markets focusing on distinct products [80]. Drugs are a primary commodity [56], [80], while ≈13% are related to hacking [21], [47]. Users include traders, customers, and administrators [61]. Cryptocurrencies like Bitcoin (BTC), Monero (XMR), or Ethereum (ETH) facilitate transactions [81]. Marketplace data includes product and trader details [43], [55], accessible post-payment [48]. Insights from marketplaces aid in understanding cybercrime markets [82], serving as early warning systems for potential data leaks [48].
- **Vendor shops**, smaller marketplaces, offer limited product variety and insights compared to larger markets [69].
- **Dark web forums** are akin to Clear web forums [47], enabling user communication [21]. Varying in size and topics, they are hierarchical and contain threads classified into categories. Posts contain textual content, hyperlinks, media, and other resources in languages like English, Chinese, and Russian [12], [17], [48]. Rich in text and metadata [12], [43], [66], forums include participant lists and user profiles with registration and last visit timestamps [12]. They offer insights into the identity, motivation, strategies, goals, and tactics (Tactics, Techniques and Procedures (TTP)) of hacking individuals [38], [43], [65], [83]. Forums also host relevant CTI-related topics [21], [24], [34], [48], [54], [82].
- **Internet Relay Chat** channels allow real-time plain text message exchange among members [17], [48]. Contents aren't usually archived, requiring active monitoring [17], [43]. Internet-Relay-Chat (IRC) channels offer insights into TTP and outbound threats from hackers [43]. For Cyber Threat Intelligence (CTI), IRCs may be promising, given hackers' freely communicated knowledge and goal sharing due to the lack of archiving in IRC channels [54].

B. POSSIBLE INFORMATION AND RELEVANCE OF HIDDEN SERVICES

This section is about what information was found in the various information sources and how the relevance ranking relates to the information found.

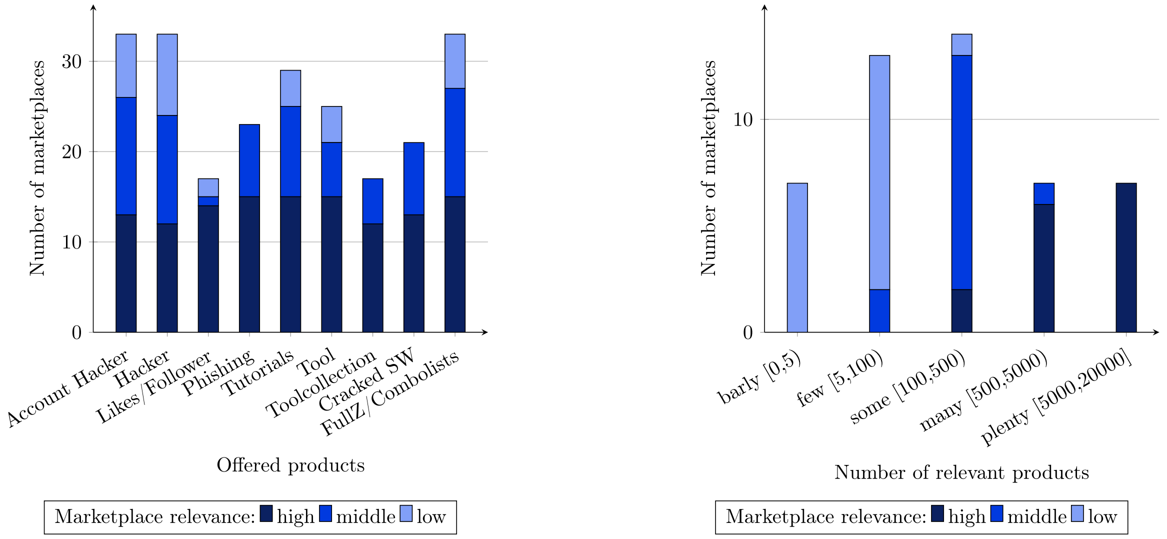
1) INFORMATION UNCOVERED IN THE DARK WEB

a: DARK WEB MARKETPLACES

In Fig. 2 we present the extracted product types on marketplaces and their total occurrences. The following categories are identified:

- **Account Hacker**: Offerings of paid services for hacking accounts.
- **Hacker Services**: Offerings of hacking for cell phones, computers, websites, grade improvements, email/SMS spamming, phishing, and DDoS. These services require interaction and don't disclose target information.
- **Social Media Enhancement Services**: Provide likes, followers, views, comments, or members for platforms like Instagram, YouTube, and Twitter, impacting disinformation campaigns, recruitment, awareness, and advertising.

[7] The darknet serves as underlying infrastructure for the dark web, similar to relation of the internet and the world wide web.



(a) This figure shows the offered product types (with a focus on IT security product) on dark web marketplaces. The relevance of a dark web marketplace is indicated by the tint of the color. The darker the color, the more relevant is a dark web marketplace.

(b) Estimation of the number of relevant products that exist in the marketplaces, divided into intervals intended to reflect the states "there are hardly any or none, few, some, many, very many relevant products in the marketplace".

FIGURE 2. **On the left is shown what is offered on marketplaces and on the right, how many products exist on the marketplaces. In both graphs 48 marketplaces are considered.**

- **Phishing and Scam Sites**: Offered with guides and tutorials for creating them.
- **Tutorials**: Cover various topics including botnet, phishing, 2FA bypass, social engineering, and penetration testing.
- **Tools and Cracked Software**: Offerings of viruses, trojans, Remote Administration Tool (RAT), ransomware builders, antivirus, anti-detection, MAC address changers, malware removal, and vulnerability scanners. Also, cracked software versions like Adobe Photoshop, WinRAR, and CyberGhost are available.
- **FullZ/Combolists**: Offerings of email lists, password lists, and personal information.

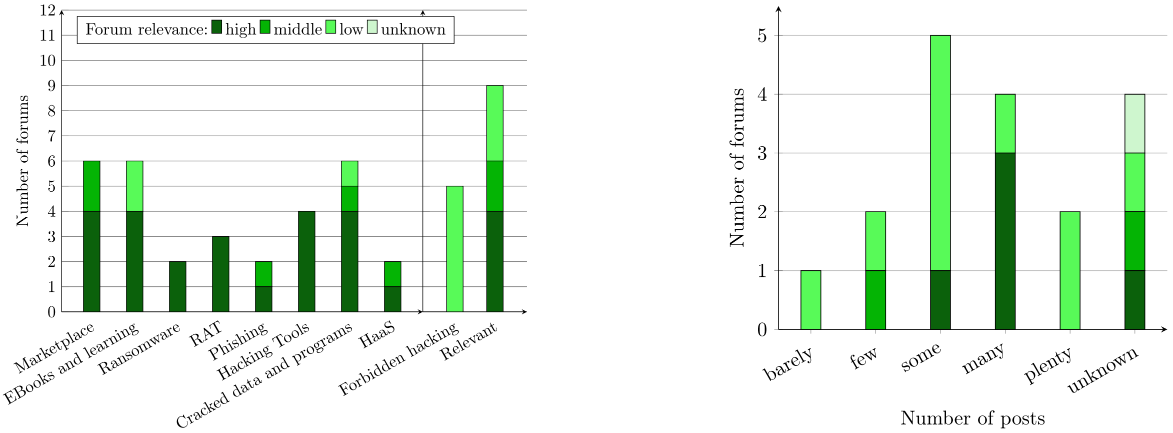
Not included in Fig. 2a are standard password lists or onion link lists due to rarity. Some markets sell invitation codes or links for other markets (Genesis Market and Benumb Market), potentially offering different products. Additionally, data extraction found CVE numbers on 7 marketplaces, often linked to the "Trillium Security MultiSploit Tool", featuring exploit generators for various CVE numbers. One such example is presented in Fig 3. The range of relevant products across the analyzed marketplaces in this paper varies between 0 and 1.7 × 10^4. In Figs. 2b to 6, an estimated upper bound of relevant products is plotted on the x-axis. This upper bound is based on the number of products within all relevant (sub)categories of the marketplace. It's important to note that not all products within these categories may be crucial (false positives), and there might be relevant products in uncounted categories (false negatives), such as misclassified items, though this is less common. Additionally, this data reflects the analysis period, and marketplace product numbers continuously change.

> *"RedLine is one of the most widely deployed information stealers that can grab Windows credentials, browser information, cryptocurrency wallets, FTP connections, banking data, and other sensitive information from the infected hosts. I want to highlight that there is 2 version of this stealer, Lite and Pro. I am selling the Full Red Line Stealer with the whole available Options, in Lite Version most advanced Options are not active and disabled. so kinda useless. be aware what you are buying. some readings."*
> REDACTED REFERENCES BY THE AUTHORS CONTAINING THE CVE-ID `cve-2022-1096`.

FIGURE 3. **Example for CVE extracted data.**

b: DARK WEB FORUMS

Parallel to marketplaces, Fig. 4 displays forum topics, post content, and total post counts, showcasing significant differences in content and size among visited forums. Approximately 5 out of 18 forums prohibit hacking or illegal content, rendering them irrelevant to this paper's research. Fig. 4a outlines forum themes, including marketplaces and forums prohibiting trading. Some forums host marketplaces for product advertisements or requests, like the CryptBB forum offering ransomware, keyloggers, RAT, hacking books, and 0-Days. Others feature account hacking services and DDoS attacks. The second category encompasses valuable sources for learning hacking skills, including programming resources, SQL injection examples, hacking challenges, and hacking books, offering the most relevant information. Forums also mention various ransomware, RAT, and hacking



(a) Most relevant information originates from higher relevant forums, with notably less from lower relevant ones. This conclusion was drawn from exploring the initial page of each interesting thread. However, this method might have missed some content, indicating that the figure does not provide a comprehensive view of the forums.

(b) In contrast to the marketplace figure, this count represents total posts, not specifically relevant ones. The study included significantly more low-relevant forums. However, post counts remain unknown when pages were inaccessible or login wasn't possible. *Barely* indicates up to 5×10^2 posts, *few* up to 1×10^3, *some* up to 1×10^4, *many* up to 3×10^4, and *plenty* up to 5×10^6 posts.

FIGURE 4. Post content and topics are shown on the left, and the total number of posts is shown on the right. Both bar charts show 18 forums.

tools like phishing tools, stealers, keyloggers, and crypters. Threads in marketplaces contain ransomware source code or Android RAT. The "cracked data/programs" section includes posts disclosing Twitter databases or hacked PayPal accounts, akin to forums where individuals offer hacking services as Hacking as a Service (HaaS). Notably, the bulk of relevant information comes from higher-relevant forums, with significantly fewer contributions from lower-relevant forums, many of which prohibited hacking content *cf.* Fig. 4b. In Fig. 4b, the post counts are displayed, albeit with some missing data due to access issues such as forum unavailability, registration problems/restrictions,[8] or payment requirements. CVE numbers were discovered in 7 forums, enabling successful CVE searches *cf.* Fig. 8a. These instances feature vulnerability news, security recommendations, and the sale of exploits. Compared to marketplaces, forums more frequently lack IT security-related information. About 47 % of forums and 13 % of marketplaces yield no relevant information, while interesting product types are almost always present on marketplaces.

c: HIDDEN SERVICES

In Fig. 5, the investigation process with final classification of 1 186 906 onion addresses is depicted. Due to issues like invalidity, duplicates, subdomains, and paths (step 1), only 5 560 addresses were automatically crawled. Of these, only 1 698 were manually classified (step 3) due to non-reachability, content duplication, and the removal of pornographic material (step 2) *cf.* §III-B. However, Step 2 did not eliminate all such pages, leading to their re-inclusion in step 3 (9 %) where content was textually saved for classification. Categorization of Onion domains involved dividing them into seven classes. Many websites, particularly in the single vendor stores category, exhibited similar content and page structure, potentially indicating phishing sites, though these were not deeply examined in Fig. 5. A significant proportion (74 %) of manually classified Hidden Services were single vendor shops. Despite low IT security relevance, the availability of hacked money accounts added interest. Another class encompassed hidden wikis, directories, and search engines, not providing direct IT security insights but serving as gateways to relevant sites. The *Miscellaneous* category (7 %) comprises various hidden services, including political pages, private individual sites, Bitcoin-related information, and a range of tools/services like Bitcoin generators, email providers, escrow, and hosting services. The *Forums, Marketplaces, and IRC* class includes sites covered in manual analysis within the present work and others not found using the entry points in this work. Challenges outlined in §IV-C emerged throughout the analysis, including linguistic barriers and blocked pages taken down by authorities. The most relevant category involved hacking-related pages (1 %). Prominent were single vendor stores offering hacking services lacking information about tactics, techniques, and targets. Noteworthy sites included an exploit builder for CVE numbers CVE-2018-0802, CVE-2017-11882, CVE-2014-1761, and CVE-2012-0158, a RAT, and a database containing leaked data.

2) RELEVANCE FACTORS INTERDEPENDENCIES

In this study, relevance relies on the quantity and nature of pertinent products/posts, as outlined in the

[8] Some forums require challenges to register, *e.g.*, proof of black-hat hacking activities.

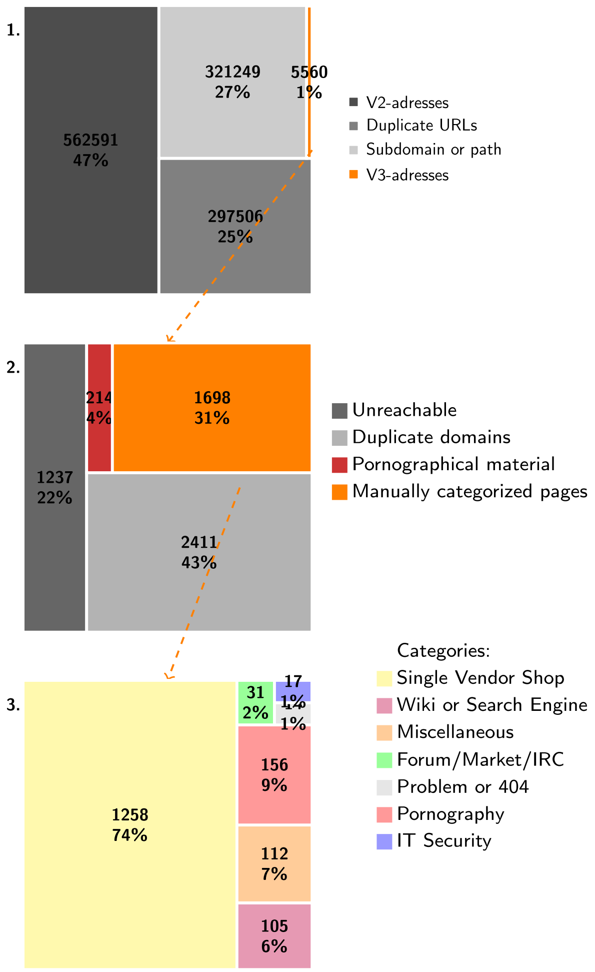


FIGURE 5. Procedure for examining hidden services. While steps 1 and 2 are automated, the last step is performed manually. In step 2 we look for duplicates based on the base domain. It should be emphasized that only 0.00143% of all given onion addresses have a verifiable, direct link to IT security.

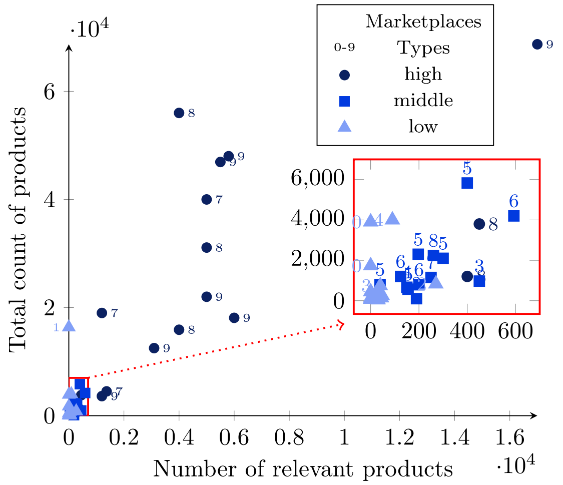


FIGURE 6. The chart illustrates relevance concerning the number of products available. Some marketplaces ignore the topic hacking, leading to lower relevance despite a high product count of up to 9 products. In this graph, outliers can be justified by the diversity of product types offered, as indicated by the label *Types*. The relevance is classified as *high*, *middle*, or *low*.

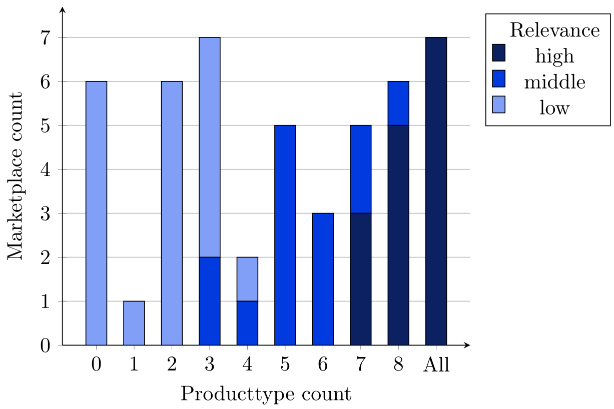


FIGURE 7. This graph shows the relevance in relation to the variability of the product types. The product types are listed in Fig. 2a and include, *e.g.*, FullZ, tools, tutorials or hacking services. It can be clearly seen that the number of product types offered has an influence on the relevance of a marketplace.

methodology *cf.* §III. This section delves into the factors affecting relevance using the gathered data.

a: MARKETPLACE RELEVANCE

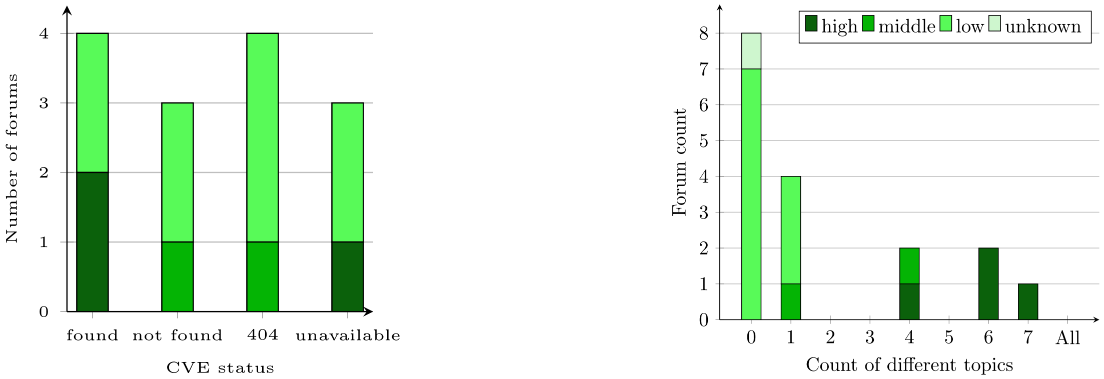
Marketplace relevance is tied to the number of potentially relevant products, depicted in Fig. 2b and Fig. 6. Total product counts often don't align with IT security relevance, with a small portion attributed to such products as seen in Fig. 6, where outliers frequently indicate restricted offerings. Usually, higher relevance accompanies a larger collection of hacking-related products. However, four outliers in Fig. 6 defy this trend: three at slightly lower relevance but with more diverse products, and one with limited offerings despite medium relevance.

b: CORRELATION WITH PRODUCT TYPES AND CVE NUMBERS

The variety of product types (*e.g.*, Account Hacker, FullZ, Tutorials, Tool) significantly influences a marketplace's relevance *cf.* Fig. 7. A wider range generally enhances relevance, yet Fig. 2b lacks a definitive correlation between offered product types and relevance. Drawing conclusive statements requires a more extensive investigation of additional marketplaces. Low-relevance markets often lack extensive tool collections, possibly due to a scarcity of hacking tools and limited inventories held by traders. Notably, hacking services, tutorials, and combo lists persist as prevailing goods across all marketplaces. Another correlation can be seen based on the finding of a CVE number. In all 7 cases the marketplace relevance was classified as high when a CVE number was found.

c: FORUM RELEVANCE

The relevance of forums heavily relies on the diversity of topics covered in the posts. Highly relevant forums tend to encompass a broader range of IT security-related



(a) Show the number of forums, which contain CVE information and their corresponding relevance, indicated by their color. The more relevant, the darker the color.

(b) Show the number of forums plotted against the count of different topics classified as security relevant. The colors of each bar depict the relevance, in terms of security, of the examined forum. The more relevant, the darker the color.

FIGURE 8. The relevance of a forum depends on the number of different topics in the forum.

themes *cf.* Fig. 8b. However, relevance doesn't directly correlate with the discovery of a CVE number, as shown in Fig. 8a. Among the 4 forums where CVE numbers were present, noticeable differences exist: high relevance forums (here 2) and low relevance forums (here 2). Further investigation based on the data might confirm varying occurrences across relevant forums. High relevance forums contained an offer of an exploit related to a CVE number, while low relevance forums contain links to clear web vulnerability descriptions. Relevance correlates with potential relevant posts and the ratio of potential relevant threads, as shown in Fig. 9. Yet, direct correlation with the total count isn't clear due to limited data. Forums surpassing 500 relevant threads and 1 000 relevant posts tend to demonstrate high relevance, while those below indicate medium or low relevance.

C. CHALLENGES OF INFORMATION GATHERING AND HOW TO DEAL WITH THEM

Information collection involves manual or automated methods, addressing two main aspects: identifying onion addresses (dark web domain addresses) and documenting the content [63]. Several challenges arise in this process, requiring resolution, acceptance, or mitigation, as elaborated in the subsequent discussion.

TABLE 3. Summary of challenges and their solutions or mitigations.

Challenge	Description	Mitigations
Fast Pace	Hidden Services are dynamic and availability is sometimes instable.	Re-crawl multiple times and keep crawlers up-to-date to get access or confirm unavailability.
Speed and Linking	The bandwidth of Tor is smaller and interpage linking is relatively poor.	Use clear web Tor directories, hidden directories, or search engines to spread the crawl seed.
Language Diversity	The examined web pages were mainly written in English or Russian.	Use a translation engine or specialized language models for analysis.
Access Restrictions	Multiple DWFs and DWMs have access restrictions to open information.	Use CAPTCHA breakers or plan with necessary interactions.
Scalability	Queuing systems of HS prolong crawling and file downloads may contain malware	Prepare malware scanners and distributed crawling.
Inconsistencies & Heterogeneity	Sites follow no strict schemata to allow easy crawling.	Specialized crawlers tailored to possibly only one site must be written.
Research/Tools	Datasets for the dark web are not published and tools are quickly outdated.	The entry fee to get going is higher than in the clear web.
Anti-Crawl Mechanisms	CAPTCHAs are a massive problem as they await everywhere.	Manual action might be necessary or further research into CAPTCHA-breakers.

a: FAST PACE

The dynamic nature of the dark web presents significant challenges in information collection [21], [22], [47], [52], [84]. Transient onion addresses, frequently undergoing changes or deactivation, exacerbate these challenges. These addresses often face seizure or go offline, further complicating consistent access. In addition, the presence of multiple addresses for single sites and intermittent downtime due to server maintenance or issues add to the complexity. The evolving structure of hidden services complicates automated crawling, necessitating regular updates to crawling scripts and increasing maintenance efforts [32].

b: SPEED AND LINKING

The Tor network, burdened by its encryption, decryption, and routing processes, operates at a slower pace compared to the conventional clearnet which presents a notable disadvantage for information extraction [17], [53]. The absence of a centralized repository for onion addresses, coupled with the

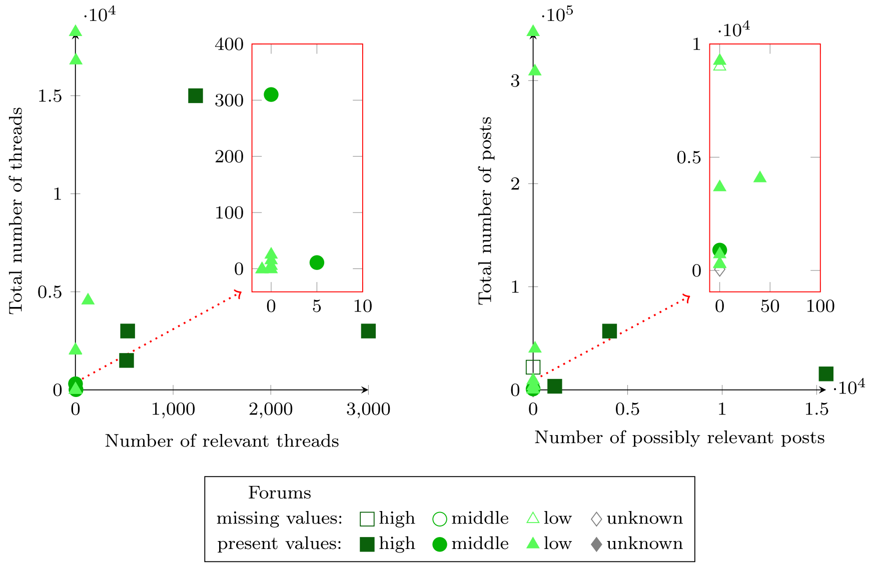


FIGURE 9. The forum's relevance hinges on the count of relevant threads (left) or posts (right). Graphs illustrate that exceeding 500 relevant threads or over 1 000 relevant posts denotes high relevance, while falling below these thresholds indicates medium or low relevance.

reluctance of many HS operators to increase their visibility, impedes the possibility of a comprehensive analysis of all HS. A substantial portion of the content remains non-linked and, therefore, inaccessible, adding to the complexity of data acquisition [12], [85]. Furthermore, redundant and looping links on the dark web pose additional challenges in automatic data extraction processes, necessitating careful consideration in their handling [21].

c: LANGUAGE DIVERSITY

In the HS of the Tor network, the most common languages we encountered where English and Russian. For automatic analysis, translation engines can be used, although the quality of these cannot currently be compared to a native human speaker [12], [49]. In addition to the variety of languages, abbreviations and code words are used [86]. Those code words make it hard to understand the actual meaning [63].

d: ACCESS RESTRICTIONS

Many marketplaces and forums have implemented access restrictions, as detailed in *cf.* Table 4. Sites often require an account to access information, including posts or product lists, necessitating registration. In certain instances, forums require payment (in 28% of the cases) or an invitation during registration. Sites demanding registration fees or vouches present legal and ethical dilemmas for researchers due to potential support of criminal activities [12], [17], [65], [69] *cf.* §III-C, leading to non-interaction in this study. Approximately ≈ 36% of accessible, understandable forums offer additional information or rights for a fee ranging from \$5 to \$250 *cf.* Fig. 10. Marketplaces, similar to those in the Clearnet, require payment for products, granting access to detailed product information. Dark web marketplaces often impose a security fee on sellers, refundable or bypassable under specific conditions, varying from \$50 to \$1 000 *cf.* Fig. 10. This seller fee explains the higher average prices, being a deposit in some cases. Fig. 10 reveals that the highest fees are associated with more relevant marketplaces/forums. In marketplaces, higher fee tiers correlate with a larger product range, whereas in forums, especially those requiring an entry fee, the exact number of posts remains unknown. In a subset of forums, specifically ≈ 24% of those analyzed, additional interaction mechanisms are employed. These interactions may include solving a challenge, writing posts, or responding to specific content. The *CryptBB* forum exemplifies a combination of these mechanisms, requiring users to solve a challenge and make three public forum posts to access more information. Moreover, responding to the original post can unveil additional hidden content. A similar approach is adopted in other dark web forums, where interaction with hidden content is facilitated. DWM frequently employ trust levels as an access control mechanism. Users can advance to higher trust levels through purchases, ratings, and secure account settings. In AlphaBay market, it was observed that new registrants at trust level 0 could not view all products; for instance, only two out of four products listed in a category were visible. Reference [78] noted in her research that Silk Road hosted secret offers, not publicly accessible but available only through URLs distributed via out-of-band mechanisms like private messages (*e.g.*between seller and buyer). Moreover, traders could operate in a 'secret mode,' with their products and information page not directly linked by SilkRoad [78]. The AlphaBay case suggests the presence of either secret products or products visible only after attaining a certain trust level.

TABLE 4. This table presents an overview of all accessible forums and marketplaces that are at least partially in English, totaling 25 forums and 48 marketplaces. The table aims to indicate the availability of information without requiring a login. It also offers insights into the frequency of successful logins, the data necessary for registration, and the factors hindering registration. These hindrances include issues like server errors, blocked IP addresses, JavaScript requirements, or suspended registrations.

Properties	Options	DNF #	DNF %	DNM #	DNM %
Openly available	No	11	≈ 44%	26	≈ 54%
	Some	6	≈ 24%	2	≈ 4%
	All	9	≈ 36%	20	≈ 42%
Registration information	Username&Password	17	100%	46	100%
	E-Mail	11	≈ 65%	5	≈ 11%
	E-Mail Validation	1	≈ 6%	0	0%
	PGP-Key	3	≈ 18%	1	≈ 2%
	PIN	0	≈ 0%	23	≈ 50%
	Login/Personal passphrase	0	≈ 0%	5	≈ 11%
	Nickname	0	≈ 0%	9	≈ 20%
Registration hurdles	Successful	10	≈ 40%	42	≈ 88%
	Not available	3	≈ 12%	1	≈ 2%
	Payment/Invitation	7	≈ 28%	0	≈ 0%
	Problem	4	≈ 16%	4	≈ 8%
	Offline (not tried)	1	≈ 4%	1	≈ 2%
Actions for information	Admin Approval	3	≈ 12%	On marketplaces, it is common to move up trust levels through interactions (purchases)	
	Payment	9	≈ 36%		
	Invitation	3	≈ 12%		
	Interaction	6	≈ 24%		

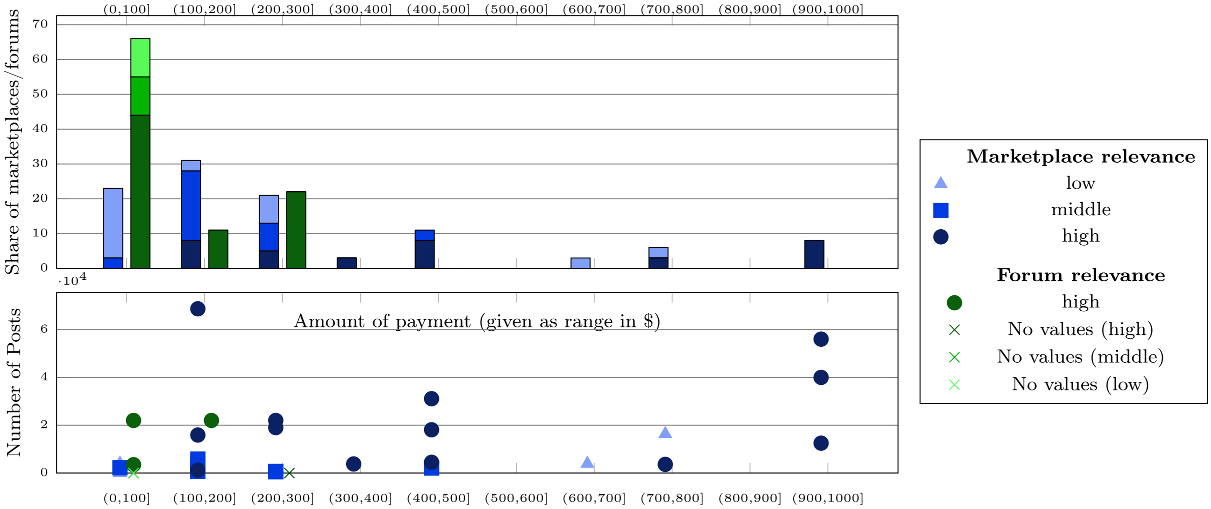


FIGURE 10. This graphic depicts the potential costs involved, encompassing all forums (9) and marketplaces (40) that require a payment. In forums, payments are predominantly for access, additional information, or extended rights, whereas in marketplaces, payments are generally for obtaining trader status, in addition to product purchases. Note that the costs of products themselves are not included in this graphic.

e: SCALABILITY

The Tor network hosts numerous websites (≈ 750 000 onion addresses[9]), with individual sites often featuring many subpages. Among the HS analyzed in this study, the largest marketplace offered around 68 700 products, while the largest forum hosted about 350 000 posts. For practical reasons a prioritized approach is essential. Large-scale data extraction requires automated crawlers capable of handling unstructured and inaccessible data in a scalable manner. Distributed crawling is one effective method [55]. Consideration of the memory load for downloaded content and efficient resource utilization is critical. Given the dynamic nature of content, especially on forums, high throughput is necessary to maintain data coverage and freshness. However, excessive requests to the same domain can overwhelm a website, inadvertently triggering a DDoS attack, an outcome to be avoided [12], [31], [49], [55]. To prevent overload, many marketplaces (approximately %ageMarketsDDoSQueue % analyzed) have implemented a queue system *cf.* Fig. 11, more frequently than forums. This security mechanism, independent of a marketplace's IT security relevance or size, ensures availability. The queue

[9] Date 2023-12-22, https://metrics.torproject.org/

mandates a waiting period before entry, prolonging both manual and automated crawling times. Crawling multimedia files significantly increases total crawling time and doubles the average file size compared to HTML files alone. Forums typically contain a mix of content types, from static files like HTML and PDF to dynamic text and multimedia data. To avoid the additional load and risks associated with multimedia files (e.g., malware or illegal content), they can be detected and annotated rather than downloaded [38].

f: INCONSISTENCIES & HETEROGENEITY

The dark web's diversity of forums and marketplaces necessitates tailored understanding and customization for automated crawling [17]. Units of measurement vary across sites; for example, seller deposits are in Euros on Cocorico Market, BTC on Babylon Market, XMR on DarkOrbit Market, and US-Dollar (USD) on others. Conversions, subject to fluctuating rates, are momentary solutions; automated processing must also account for varied spellings of the same values (*e.g.*, "50 USD" vs "$50") [80]. Posts and products are often inconsistently categorized, misplaced in wrong or multiple categories, or found as empty or duplicate listings. Users may also find category assignments unclear, further complicating data analysis [61], [87], [88].

g: RESEARCH/TOOLS

A comprehensive and in-depth analysis of the dark web remains largely underdeveloped. Consequently, there is a scarcity of supportive tools for dark web investigation, and those that exist are often not publicly available, primarily due to concerns over their dual-use potential. For instance, [38] withheld their dataset from publication to prevent misuse. Furthermore, the dark web's rapid evolution renders datasets quickly outdated and obsolete, often covering only a limited number of websites [63], [89].

h: ANTI-CRAWL MECHANISMS

Information extraction from the dark web is further complicated by anti-crawling mechanisms and DDoS prevention strategies, mainly implemented in CAPTCHAs. Some websites have systems to detect excessive requests from the same Tor identity, leading to blocking of that identity, necessitating a change of identity for continued access. To circumvent these mechanisms, automated crawlers can mimic human behavior, such as slowing crawl speeds or introducing random delays between page requests [38]. However, the challenge persists even in manual investigations, requiring the ability to automatically detect and reset the identity upon detection. Additionally, websites may employ CAPTCHA challenges or redirect to a DDoS-CAPTCHA page after a certain number of page visits, as noted by [57] in the context of many marketplaces. These mechanisms extend the execution time for both manual and automated crawling and necessitate additional considerations in the crawling script or the possibility of recall due to incorrect data display [43], [49]. Reference [87] also highlight an anti-crawling technique that randomizes the HTML/CSS of login pages to further impede automated data collection. The most significant anti-crawling challenge encountered in this study is the use of CAPTCHA. According to [90] and [91], CAPTCHA can be classified into four types: text-based, image-based, video-based, and audio-based. In this research, CAPTCHA were encountered in various situations, including i) DDoS prevention, ii) login processes, iii) registration procedures, iv) post writing, v) handling of numerous emergent requests, and vi) marketplace browsing.

Fig. 11 presents the frequencies of DDoS, registration, and login CAPTCHA. Notably absent were video- and audio-based CAPTCHAs. Fig. 12 presents CAPTCHA examples, we encountered during our analysis, ranging from clock CAPTCHAs *cf.* Fig. 12a to chess-like CAPTCHAs *cf.* Fig. 12d. It was observed that forums were less likely to implement CAPTCHA mechanisms compared to marketplaces, and the complexity of these mechanisms was also generally lower in forums. This disparity could be attributed to the higher proportion of low-relevance forums studied compared to low-relevance marketplaces. Marketplaces also have a higher likelihood of being targeted for DDoS attacks due to their significant role in illicit activities and their potential impact on criminal trafficking. The types and difficulty levels of the individual CAPTCHA encountered during this research are detailed, providing valuable insights for future studies on CAPTCHA complexity. It was observed that CAPTCHA appeared more frequently during registration processes in forums, in contrast to marketplaces where CAPTCHA were commonly used at both registration and login stages. The tendency to employ CAPTCHA mechanisms varied with the relevance of the HS; higher-relevance sites were more inclined to use them. This could be partly due to the larger size of the marketplaces, as they might aim to enhance their security measures. In forums, the likelihood of CAPTCHA implementation increased with relevance, possibly reflecting the advanced IT security knowledge of the operators. The content focus within forums also differed based on their relevance; low-relevance forums generally covered a broader range of topics and frequently restricted hacking-related discussions, while high-relevance forums often included hacking-specific categories. Consequently, the need for robust security measures and the risk of unauthorized access were greater in forums of higher relevance.

For automated crawling of CAPTCHA-protected pages, the CAPTCHA must be either solved, bypassed, or the pages ignored [50]. There are three primary methods for tackling CAPTCHA [92]: i) halting the crawling process at a CAPTCHA and requiring manual resolution [38], [54], [56], [59], [80], ii) utilizing external services that employ human solvers for CAPTCHA, such as *Anti-Captcha*[10] and *2Captcha*[11] [87], [93], and iii) developing automated

[10] https://anti-captcha.com/de

[11] https://2captcha.com/

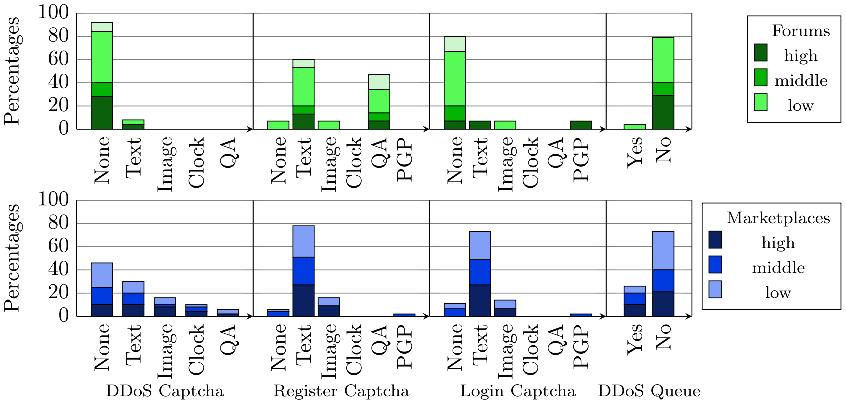
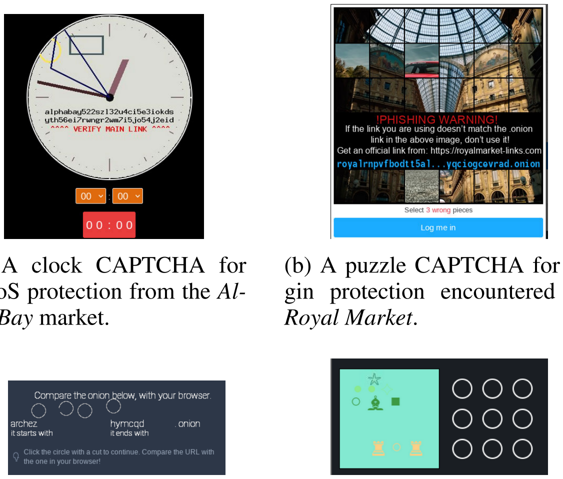


FIGURE 11. It can be seen that forums have implemented far fewer DDoS captchas than markets. At the same time, marketplaces have more difficult DDoS captchas. In the forums examined, captchas often only occur when registering and not when logging in, while both are often implemented on marketplaces. Furthermore, it can be seen that a DDoS queue is implemented far more frequently on marketplaces than on forums.


alphabay522szl32u4ci5e3iokds
yth56ei7rwngr2wm7i5jo54j2eid
^^^^ VERIFY MAIN LINK ^^^^
00
00
00 : 00

!PHISHING WARNING!
If the link you are using doesn't match the .onion link in the above image, don't use it!
Get an official link from: https://royalmarket-links.com
royalrnpvfbodtt5al...yqciogcevrad.onion

Select 3 wrong pieces
Log me in

(a) A clock CAPTCHA for DDoS protection from the *AlphaBay* market.

(b) A puzzle CAPTCHA for login protection encountered on *Royal Market*.

Compare the onion below, with your browser.

Click the circle with a cut to continue. Compare the URL with the one in your browser!

(c) A puzzle CAPTCHA for registration protection encountered on *ArcheTyp*.

(d) A chess-like CAPTCHA for DDoS protection encountered on *ViceCity*.

FIGURE 12. Example CAPTCHAs encountered during the study.

CAPTCHA solvers to circumvent service costs [60], [90], [92], [94]. Fig. 13 and Fig. 14 display the challenges encountered during data extraction from DWM and DWF, in the order experienced during a visit. These figures clarify the necessity of solving CAPTCHA to access IT security information. The depicted visit process for a HS encompasses all HS that are accessible and visible without JavaScript and have not been seized. A potential initial barrier is the DDoS-CAPTCHA encountered upon page entry. If this CAPTCHA exists and remains unsolved, no IT security information can be accessed. The availability of information upon entering the page is then assessed. If all information is readily available, no additional efforts are required. However, if only partial or no information is available, registration, possibly involving another CAPTCHA mechanism, is needed. This ensures restricted information disclosure without CAPTCHA resolution. Subsequently, solving a login CAPTCHA might be required.

In the marketplace analysis presented in Fig. 13, 48 marketplaces were examined from which 13 medium and low relevance marketplaces could be accessed without resolving any CAPTCHA. All necessary information can be accessed without solving any CAPTCHA from 8 out of 28 forums, although these forums typically contain limited IT securityrelevant information (*cf.* Fig. 14). Partial information can be extracted from 5 forums, which vary in their relevance.

V. DISCUSSION

The discussion in §V-A initiates with an inquiry into the methodologies for uncovering IT security-related information on the dark web. This involves a comparative analysis of manual versus automated crawling techniques. Key to this comparison is an examination of the challenges associated with each extraction method. The subsequent segment of the discussion evaluates the dark web's potential as an open information resource. In this context, research questions and are explored in greater depth. This section concludes with the limitations encountered in this study as well as future work *cf.* §V-C. It outlines the various decisions in the research that imposed constraints. Furthermore, the rationale behind these limitations and their consequent implications are thoroughly discussed.

A. MANUAL VS. AUTOMATED CRAWLING

We identified, that IT security-relevant information can be extracted from the dark web through manual or automated crawling methods, which answers **RQ1**. Each approach comes with its own set of pros and cons as identified by the several problems outlined in §IV-C. Manual extraction, often criticized for being infeasible and time-consuming, is inefficient and prone to errors [26], [95]. It becomes impractical when dealing with large volumes of data. Additionally, manual extraction may be undesirable due to

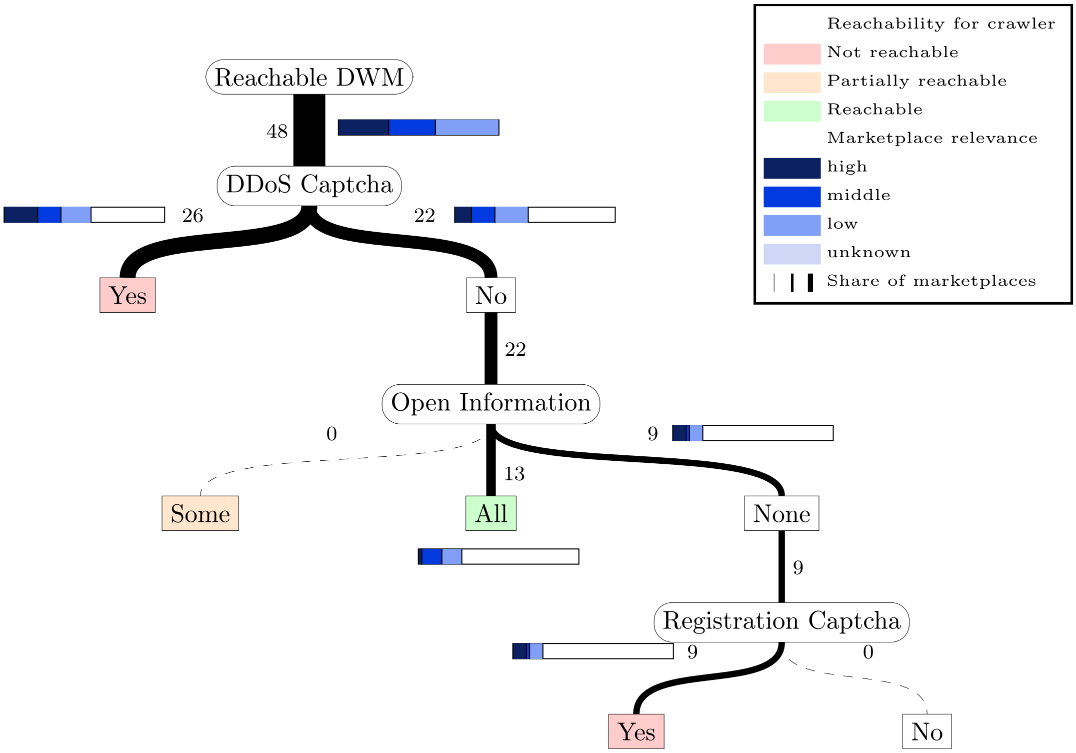


FIGURE 13. This figure illustrates the historical pattern of DWM visits. It depicts whether relevant IT security information can be extracted from DWM without the need to solve CAPTCHA. The figure effectively shows the sequence of a typical DWM visit. It reveals that in most marketplaces, solving a CAPTCHA is a prerequisite for accessing IT security-related information.

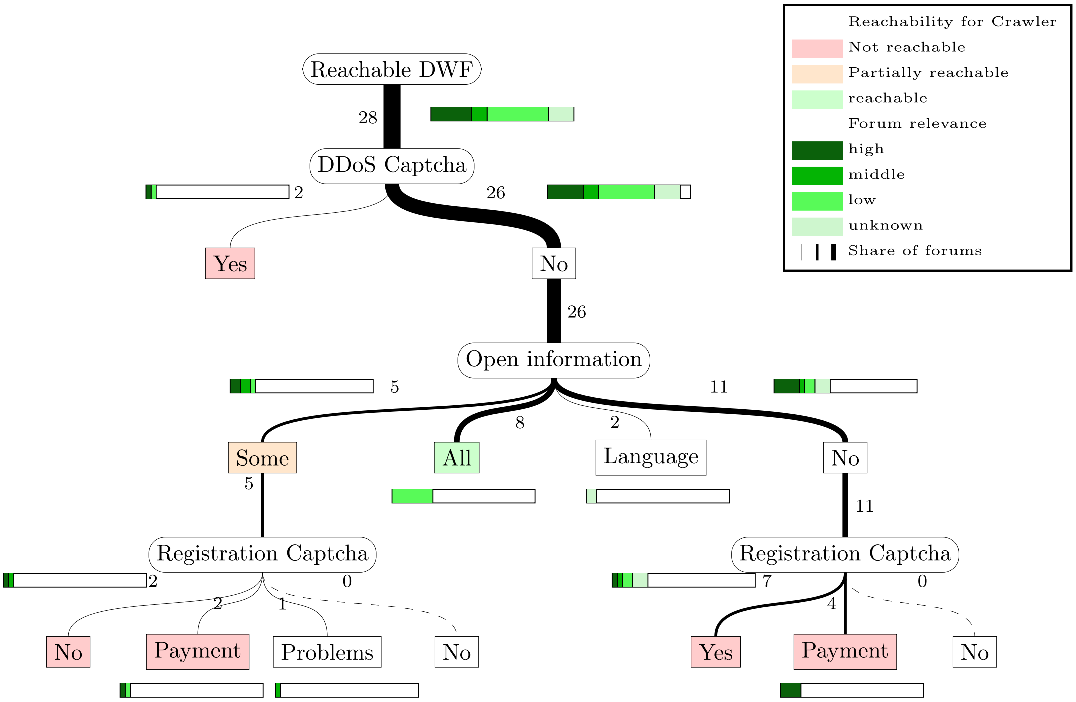


FIGURE 14. This figure illustrates the history of DWF visits. It demonstrates, based on the progression of a DWF visit, whether relevant information can be extracted from DWF without the need to solve CAPTCHA. The figure indicates that complete information extraction from a forum is typically possible only in low-relevance forums.

personal protection reasons, such as avoiding exposure to pornographic content. The process is significantly slower for humans, especially in cases requiring frequent revisits to the same page for monitoring purposes. Ensuring traceability, reproducibility, and facilitating analysis during manual extraction also adds to the time burden. Errors during manual extraction are inevitable due to human factors such as lack of concentration. However, manual extraction is indispensable in certain scenarios. Understanding the structure of forums and marketplaces is crucial for accurate data storage, necessitating manual inspection of the page and analysis of the Document Object Model (DOM) element to adjust the page parser accordingly *cf.* §IV-C. Manual crawling is also essential for gaining a comprehensive understanding of the subject and obtaining valuable insights. As highlighted in §IV-C, automated crawling in the dark web is fraught with challenges, answering **RQ2**. A significant obstacle for automated crawling is overcoming anti-crawling mechanisms, as evident in Fig. 14 and Fig. 13. Many forums and marketplaces necessitate solving a CAPTCHA to access comprehensive information.

Currently, researchers often resort to manually solving CAPTCHAs, necessitating continuous human involvement. Only a limited number of studies have focused on the automated resolution of these CAPTCHAs [60], [90], [94]. Therefore, human intervention or a fully manual crawl is required to gather all pertinent information.

B. DARK WEB AS OPEN INFORMATION SOURCE

The primary objective of this study is to utilize the dark web as a OSINT source for IT security related information, as outlined in §I. §II presents previous research using the dark web for IT security information is reviewed. This includes extracting new cyber threats, analyzing trends, examining relationships between hidden services, classifying attachments, predicting cyber threats, and generating alerts, answering **RQ3**.

This research substantiates the notion that the dark web is a valuable OSINT resource, demonstrating various use cases. For instance, [28] focuses on generating cyber threat warnings using Twitter and the dark web. The findings, illustrated in Table 5 and Fig. 15, reveal a lower frequency of warnings based on dark web posts compared to Twitter. Their data sources included 200 dark web and deep web hacking forums and Twitter posts from 69 international security researchers and analysts.

During this study, relevant information was found on marketplaces, forums, and other HS. The investigation revealed a scarcity of IT security-related information: approximately 88% of marketplaces and 53% of forums contained relevant data. The findings encompass hacking tools, service offerings, cracked software, and leaked data (*cf.* §IV-B1 and **RQ4**). However, a significant amount of information remained inaccessible due to hidden content, payment requirements, or memberships *cf.* §IV-C. Thus, more information likely remains concealed on the dark web. The semi-automatic analysis indicated that only 0.00143% of all analyzed onion addresses had a direct, identifiable link to hacking, amounting to 17 HS. [28] suggests that the clear web may yield more information and is potentially easier to access, the cost and effort of using the dark web as an information source must be taken into consideration, which is inline with [96]. It states that Twitter is primarily useful to advertise services and comment on cyber threats rather than patch prioritization. However, this inference must be approached with caution, as the clear web was not evaluated as an information source in this study, thus precluding a direct comparison. Moreover, the quantity of information in the clear web might be outweight by quality in the dark web.

TABLE 5. Cyber threats for which alerts were generated in the paper by [28] and related tweets and dark web mentions.

			Mentions	
Discovered Word	**Threat type**	**Warnings**	**Twitter/X**	**Dark Web**
mirai	Weak Spot	94	537	85
teamxrat	Ransomware	13	30	0
luabot	Trojans	12	27	0
cryptoluck	Ransomware	12	29	0
clixsense	data breach	12	26	5
gooligan	Malware	9	26	0
usbee	Malware	9	18	0
adultfriendfinder	Data breach	9	23	0
starhub	Data breach	9	82	0
badepilogue	Malware	8	21	0
evony	Data breach	8	25	0

C. LIMITATIONS AND FUTURE WORK

Initially, the research was confined to the Tor network within the dark web. The list of DWFs and DWMs scrutinized was not exhaustive, comprising only those found in public directories. Despite consulting multiple directories, it cannot be assumed that these directories encompass all existing sites. Due to the fact, that the dark web is a highly dynamic domain, with domains disappearing once in a while, our results can only be seen as a snapshot, which is backed up by findings in related work [1], [21], [28]. In addition, every research, especially manual investigations, inherits some sort of bias of the researchers, which we aimed to mitigate by our chosen and clearly described and strictly applied methodology *cf.* §III. While the semi-automated analysis identified some additional forums and marketplaces, these were not subsequently examined in detail. Due to complexity constraints, the study did not delve into all posts and products on the forums and marketplaces, focusing instead on just the first page of each. Further, a semi-automated analysis, using a handwritten crawler might have provided additional insights. However, such an attempt would likely produce an overwhelming amount of data to analyze, for our scope and was out of scope for the present work.

The absence of direct interaction also represents a limitation of this research. In summary, there was no interaction with traders. Additionally, the study did not

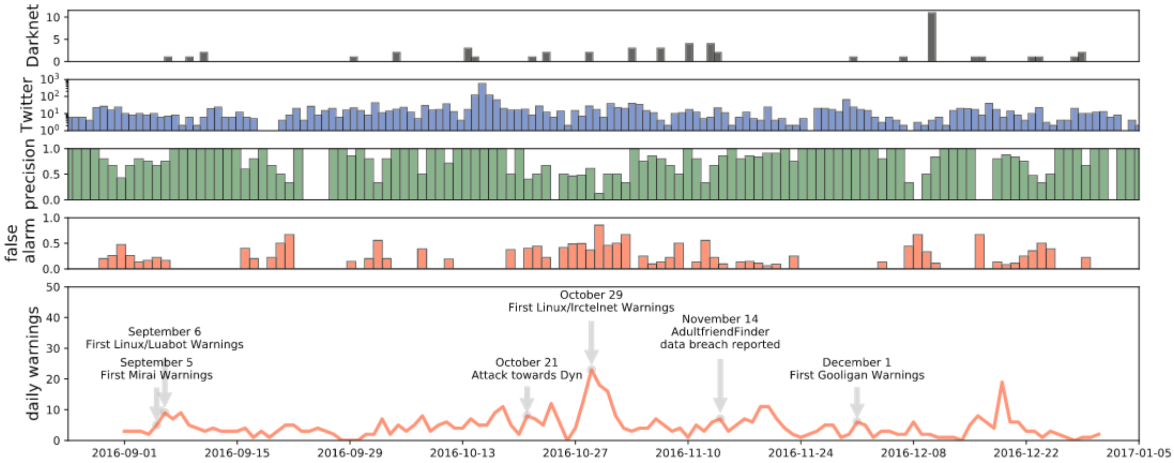


FIGURE 15. Number of daily generated warnings and the number of related tweets and dark web posts (Source: [28]).

involve responding to hidden content, challenges, or posts in forums, nor were any payments made for memberships or access to information.

A significant challenge in automatic data extraction is overcoming anti-crawling measures, particularly CAPTCHAs. For efficient and fully automated extraction, the development of a robust CAPTCHA-breaker, capable of adapting to the dynamic nature of the dark web, is essential. The current state of the CAPTCHA variants used can be instrumental in future developments. These limitations imply that the scope of information captured in this research is not comprehensive, thereby limiting its ability to represent the full spectrum of IT security information available on the dark web. Another limitation pertains to the exclusion of information from non-German and non-English Hidden Services. Had there been proficiency in languages such as Russian or French, or if automatic translation tools were employed, a broader range of information might have been available for analysis.

1) FUTURE WORK

The manual analysis in this study did not encompass the complete extraction and analysis of information from forums and marketplaces, mainly due to the involved efforts. Follow-up studies could monitor open dark web information sources for more IT security-related information tackling and mitigating the identified challenges in an automated manner, including the identification of key users and connections via social network analysis (SNA), and analysis of shared attachments, and source code could yield a more comprehensive and current overview. Forums focusing on hacking and programming are of particular interest in this context. An automated and thorough analysis of IT security-related content in these forums could offer significant insights. Moreover, leveraging a larger and more accurate dataset of the dark web, automated analyses utilizing ML could be conducted, providing additional valuable CTI information.

In the semi-automated analysis phase, onion addresses were initially extracted from a database using keywords. This approach presents upward potential with improved domain filtering from the original data. Various ML/Natural Language Processing (NLP) techniques could be employed to enhance the results. While the results of the present work indicate, that complementing CTI with automated dark web intelligence might be too costly, addressing the question of which information source – dark web or clear web – is more promising, a direct comparative study of these two sources could provide valuable insights. Future research could build upon the literature review and the comprehensive analysis of IT security-related content conducted in this study. The insights gained from this work offer a foundational overview of prior research and a current understanding of the dark web, serving as a basis for all future CTI research focused on the dark web. With a potential availability of Quantum Computing (QC) and Post-Quantum Cryptography (PQC) the current dark web may tremble, as it is based on public key cryptography, which can be broken with [97].

VI. CONCLUSION

The dark web is another open information source for CTI [1], [20], [21], [63]. This work provides an overview of existing CTI research and potential CTI-relevant information on the dark web, serving as a foundation for future researchers in designing their investigations. A key goal of this research was to analyze the information available on the dark web and assess its relevance to IT security. The distinction between this work and existing literature lies in the methodology employed and the focus of the information presented [1], [20], [21], [63]. Achieved through manual and semi-automated IR, this study identified that proactive CTI in the form of exploits, malware, and their stakeholders can be sourced from the dark web, whereas others mainly pick one specific aspect to look out for, *e.g.*, Internet of Things (IoT)-based CTI [1]. Leaked company data and cracked software, indicative of vulnerabilities and reactive information, can help protect against further data loss. Contrary to expectations, few visible

TTPs or attack targets were discovered in forums. Techniques were mainly found in tutorials and e-books, with ransomware, keyloggers, other malware, and hacking literature being significant finds. Hacking as a service and DDoS as a service offerings were noted, relevant for general security but less so for effective countermeasures unless attribution is involved. Manual investigation of 65 dark web forums, 7 single-vendor shops, and 73 dark web marketplaces revealed pertinent CTI information, while the semi-automated analysis of numerous Onion domains was less fruitful. Despite keyword-based domain extraction, the majority of HS were unrelated to hacking, highlighting the low percentage of IT security-related HS among all HS. The study identified several challenges in analysis and crawling, such as the transient nature and inaccessibility of content, the speed of the Tor network, language diversity, inconsistent categorizations, and scalability issues. Access restrictions like interactions or payments also limit the research scope. Notably, hidden content in forums, accessible via post responses, and account creation pose significant hurdles. Future research will grapple with anti-crawling measures, particularly CAPTCHA at various stages of HS access. This research illuminates new potential CTI information sources. Considering the extraction challenges and information obtained from the dark web versus the clear web, as shown in [28]'s work, the clear web be the primary information source. It can be complemented by dark web collection if (i) current CTI capabilities are not saturated, (ii) clear web data only hints at deeper threats, or (iii) the information is so niche and sensitive that its discussion might be illegal, as dark web collection is more complex. However, valuable insights can still be derived from the dark web, useful for both reactive and proactive purposes. Given the escalating cyber threats, the dark web's relevance as an information source is likely to grow in the future.

ACKNOWLEDGMENT

The authors would like to thank the members of the PANDA project for providing the dataset and the helpful discussions.

REFERENCES

[1] P. Koloveas, T. Chantzios, S. Alevizopoulou, S. Skiadopoulos, and C. Tryfonopoulos, "InTIME: A machine learning-based framework for gathering and leveraging web data to cyber-threat intelligence," *Electronics*, vol. 10, no. 7, p. 818, Mar. 2021.

[2] A. Fleck. (2024). *Infographic: Cybercrime Expected to Skyrocket in Coming Years*. [Online]. Available: https://www.statista.com/chart/28878/expected-cost-of-cybercrime-until-2027

[3] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1197–1227, 2nd Quart., 2016.

[4] F. Dong, S. Yuan, H. Ou, and L. Liu, "New cyber threat discovery from darknet marketplaces," in *Proc. IEEE Conf. Big Data Anal. (ICBDA)*, Nov. 2018, pp. 62–67.

[5] P. Kuehn, M. Bayer, M. Wendelborn, and C. Reuter, "OVANA: An approach to analyze and improve the information quality of vulnerability databases," in *Proc. 16th Int. Conf. Availability, Rel. Secur.*, Aug. 2021, p. 11.

[6] P. Kühn, D. N. Relke, and C. Reuter, "Common vulnerability scoring system prediction based on open source intelligence information sources," *Comput. Secur.*, vol. 131, Aug. 2023, Art. no. 103286.

[7] M. Bayer, P. Kuehn, R. Shanehsaz, and C. Reuter, "CySecBERT : A domain-adapted language model for the cybersecurity domain," *ACM Trans. Privacy Secur.*, vol. 27, no. 2, pp. 1–20, May 2024.

[8] M. Bayer, T. Frey, and C. Reuter, "Multi-level fine-tuning, data augmentation, and few-shot learning for specialized cyber threat intelligence," *Comput. Secur.*, vol. 134, Nov. 2023, Art. no. 103430.

[9] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Reading, MA, USA: Addison-Wesley, 2009.

[10] I. Hernández, C. R. Rivero, and D. Ruiz, "Deep web crawling: A survey," *World Wide Web*, vol. 22, no. 4, pp. 1577–1610, Jul. 2019.

[11] R. Basheer and B. Alkhatib, "Threats from the dark: A review over dark web investigation research for cyber threat intelligence," *J. Comput. Netw. Commun.*, vol. 2021, pp. 1–21, Dec. 2021.

[12] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the dark web for cyber security information," in *Proc. 11th Int. Conf. Cyber Conflict (CyCon)*, May 2019, pp. 1–21.

[13] E. Marin, M. Almukaynizi, S. Sarkar, E. Nunes, J. Shakarian, and P. Shakarian, *Exploring Malicious Hacker Communities: Toward Proactive Cyber-Defense*. Cambridge, U.K.: Cambridge Univ. Press, 2021.

[14] D. Harris, Ed., *Engineering Psychology and Cognitive Ergonomics: 12th International Conference, EPCE 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings* (Lecture Notes in Computer Science), vol. 9174. Cham, Switzerland: Springer, 2015, doi: 10.1007/978-3-319-20373-7.

[15] J. Pastor-Galindo, P. Nespoli, F. G. Mármol, and G. M. Pérez, "The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends," *IEEE Access*, vol. 8, pp. 10282–10304, 2020.

[16] J. Bergman and O. B. Popov, "Exploring dark web crawlers: A systematic literature review of dark web crawlers and their implementation," *IEEE Access*, vol. 11, pp. 35914–35933, 2023.

[17] V. Benjamin, J. S. Valacich, and H. Chen, "DICE-E: A framework for conducting darknet identification, collection, evaluation with ethics," *MIS Quart.*, vol. 43, no. 1, pp. 1–22, Jan. 2019.

[18] J. Hughes, S. Pastrana, A. Hutchings, S. Afroz, S. Samtani, W. Li, and E. Santana Marin, "The art of cybercrime community research," *ACM Comput. Surv.*, vol. 56, no. 6, pp. 155:1–155:26, Jun. 2024.

[19] S. Sobhan, T. Williams, M. J. H. Faruk, J. Rodriguez, M. Tasnim, E. Mathew, J. Wright, and H. Shahriar, "A review of dark web: Trends and future directions," in *Proc. IEEE 46th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jun. 2022, pp. 1780–1785.

[20] B. Ampel, S. Samtani, H. Zhu, S. Ullman, and H. Chen, "Labeling hacker exploits for proactive cyber threat intelligence: A deep transfer learning approach," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2020, pp. 1–6.

[21] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, and P. Shakarian, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 7–12.

[22] O. Cherqi, G. Mezzour, M. Ghogho, and M. El. Koutbi, "Analysis of hacking related trade in the darkweb," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 79–84.

[23] I. Deliu, C. Leichter, and K. Franke, "Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2017, pp. 3648–3656.

[24] M. Kadoguchi, S. Hayashi, M. Hashimoto, and A. Otsuka, "Exploring the dark web for cyber threat intelligence using machine leaning," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 200–202.

[25] P. Koloveas, T. Chantzios, C. Tryfonopoulos, and S. Skiadopoulos, "A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence," in *Proc. IEEE World Congr. Services (SERVICES)*, Jul. 2019, pp. 3–8.

[26] I. Deliu, C. Leichter, and K. Franke, "Collecting cyber threat intelligence from hacker forums via a two-stage, hybrid process using support vector machines and latent Dirichlet allocation," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 5008–5013.

[27] E. Marin, M. Almukaynizi, E. Nunes, and P. Shakarian, "Community finding of malware and exploit vendors on darkweb marketplaces," in *Proc. 1st Int. Conf. Data Intell. Secur. (ICDIS)*, Apr. 2018, pp. 81–84.

[28] A. Sapienza, A. Bessi, S. Damodaran, P. Shakarian, K. Lerman, and E. Ferrara, "Early warnings of cyber threats in online discussions," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 667–674.

[29] M. Almukaynizi, A. Grimm, E. Nunes, J. Shakarian, and P. Shakarian, ''Predicting cyber threats through hacker social networks in darkweb and deepweb forums,'' in *Proc. Int. Conf. Comput. Social Sci. Soc. Americas*, Oct. 2017, pp. 1–7.

[30] S. Sarkar, M. Almukaynizi, J. Shakarian, and P. Shakarian, ''Predicting enterprise cyber incidents using social network analysis on dark web hacker forums,'' *Cyber Defense Rev.*, vol. 1, pp. 87–102, Jan. 2018.

[31] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, ''Surfacing collaborated networks in dark web to find illicit and criminal content,'' in *Proc. IEEE Conf. Intell. Secur. Informat. (ISI)*, Sep. 2016, pp. 109–114.

[32] D. Hayes, F. Cappa, and J. Cardon, ''A framework for more effective dark web marketplace investigations,'' *Information*, vol. 9, no. 8, p. 186, Jul. 2018.

[33] R. Williams, S. Samtani, M. Patton, and H. Chen, ''Incremental hacker forum exploit collection and classification for proactive cyber threat intelligence: An exploratory study,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 94–99.

[34] S. Samtani, R. Chinn, H. Chen, and J. F. Nunamaker, ''Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence,'' *J. Manage. Inf. Syst.*, vol. 34, no. 4, pp. 1023–1053, Oct. 2017.

[35] J. Grisham, S. Samtani, M. Patton, and H. Chen, ''Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence,'' in *Proc. IEEE Int. Conf. Intell. Security Informat. (ISI)*, Jul. 2017, pp. 13–18.

[36] G.-Y. Shin, Y. Jang, D.-W. Kim, S. Park, A.-R. Park, Y. Kim, and M.-M. Han, ''Dark side of the web: Dark web classification based on TextCNN and topic modeling weight,'' *IEEE Access*, vol. 12, pp. 36361–36371, 2024.

[37] V. Adewopo, B. Gonen, N. Elsayed, M. Ozer, and Z. S. Elsayed, ''Deep learning algorithm for threat detection in hackers forum (deep web),'' 2022, *arXiv:2202.01448*.

[38] S. Pastrana, D. R. Thomas, A. Hutchings, and R. Clayton, ''CrimeBB: Enabling cybercrime research on underground forums at scale,'' in *Proc. World Wide Web Conf. World Wide Web (WWW)*. New York, NY, USA: ACM Press, 2018, pp. 1845–1854.

[39] N. Deguara, J. Arshad, A. Paracha, and M. A. Azad, ''Threat miner—A text analysis engine for threat identification using dark web data,'' in *Proc. IEEE Int. Conf. Big Data*, Dec. 2022, pp. 3043–3052.

[40] N. Tavabi, P. Goyal, M. Almukaynizi, P. Shakarian, and K. Lerman, ''DarkEmbed: Exploit prediction with neural language models,'' in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI/IAAI/EAAI)*. Washington, DC, USA: AAAI Press, 2018, pp. 7849–7854.

[41] S. Nikoletos and P. Raftopoulou, ''Employing social network analysis to dark web communities,'' in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Jul. 2022, pp. 311–316.

[42] M. Ebrahimi, M. Surdeanu, S. Samtani, and H. Chen, ''Detecting cyber threats in non-english dark net markets: A cross-lingual transfer learning approach,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 85–90.

[43] P.-Y. Du, N. Zhang, M. Ebrahimi, S. Samtani, B. Lazarine, N. Arnold, R. Dunn, S. Suntwal, G. Angeles, R. Schweitzer, and H. Chen, ''Identifying, collecting, and presenting hacker community data: Forums, IRC, carding shops, and DNMs,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 70–75.

[44] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, ''Evolution of dark web threat analysis and detection: A systematic approach,'' *IEEE Access*, vol. 8, pp. 171796–171819, 2020.

[45] E. Marin, M. Almukaynizi, and P. Shakarian, ''Reasoning about future cyber-attacks through socio-technical hacking information,'' in *Proc. IEEE 31st Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2019, pp. 157–164.

[46] M. Almukaynizi, E. Nunes, K. Dharaiya, M. Senguttuvan, J. Shakarian, and P. Shakarian, ''Proactive identification of exploits in the wild through vulnerability mentions online,'' in *Proc. Int. Conf. Cyber Conflict (CyCon)*, Nov. 2017, pp. 82–88.

[47] J. Robertson, A. Diab, E. Marin, E. Nunes, V. Paliath, J. Shakarian, and P. Shakarian, ''Darknet mining and game theory for enhanced cyber threat intelligence,'' *Cyber Defense Rev.*, vol. 1, no. 2, pp. 95–122, 2016. [Online]. Available: https://www.jstor.org/stable/26267362

[48] S. Samtani, Y. Chai, and H. Chen, ''Linking exploits from the dark web to known vulnerabilities for proactive cyber threat intelligence: An attention-based deep structured semantic model,'' *MIS Quart.*, vol. 46, no. 2, pp. 911–946, May 2022. [Online]. Available: https://aisel.aisnet.org/misq/vol46/iss2/11

[49] T. Fu, A. Abbasi, and H. Chen, ''A focused crawler for dark web forums,'' *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 6, pp. 1213–1231, Jun. 2010.

[50] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. de Paz, ''Classifying illegal activities on tor network based on web textual contents,'' in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017, pp. 35–43.

[51] Y. Yannikos, J. Heeger, and M. Steinebach, ''Data acquisition on a large darknet marketplace,'' in *Proc. 17th Int. Conf. Availability, Rel. Secur.*, Aug. 2022, pp. 1–6.

[52] C. Iliou, G. Kalpakis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, ''Hybrid focused crawling on the surface and the dark web,'' *EURASIP J. Inf. Secur.*, vol. 2017, no. 1, p. 11, Dec. 2017.

[53] J. Park, H. Mun, and Y. Lee, ''Improving tor hidden service crawler performance,'' in *Proc. IEEE Conf. Dependable Secure Comput. (DSC)*, Dec. 2018, pp. 1–8.

[54] S. Samtani, W. Li, V. Benjamin, and H. Chen, ''Informing cyber threat intelligence through dark web situational awareness: The AZSecure hacker assets portal,'' *Digit. Threats, Res. Pract.*, vol. 2, no. 4, pp. 1–10, Dec. 2021.

[55] B. Alkhatib and R. S. Basheer, ''Mining the dark web: A novel approach for placing a dark website under investigation,'' *Int. J. Modern Educ. Comput. Sci.*, vol. 11, no. 10, pp. 1–13, Oct. 2019.

[56] A. Baravalle, M. S. Lopez, and S. W. Lee, ''Mining the dark web: Drugs and fake ids,'' in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 350–356.

[57] Y. Wu, F. Zhao, X. Chen, P. Skums, E. L. Sevigny, D. Maimon, M. Ouellet, M. H. Swahn, S. M. Strasser, M. J. Feizollahi, Y. Zhang, and G. Sekhon, ''Python scrapers for scraping cryptomarkets on tor,'' in *Security, Privacy, and Anonymity in Computation, Communication, and Storage*, vol. 11611, G. Wang, J. Feng, M. Z. A. Bhuiyan, and R. Lu, Eds. Cham, Switzerland: Springer, 2019, pp. 244–260, doi: 10.1007/978-3-030-24907-6_19.

[58] Y. Xu, G. Chen, J. Wu, W. Xu, and Q. Liu, ''Research on dark web monitoring crawler based on TOR,'' in *Proc. IEEE 2nd Int. Conf. Inf. Technol., Big Data Artif. Intell. (ICIBA)*, vol. 2, Dec. 2021, pp. 197–202.

[59] M. Ebrahimi, J. F. Nunamaker, and H. Chen, ''Semi-supervised cyber threat identification in dark net markets: A transductive and deep learning approach,'' *J. Manage. Inf. Syst.*, vol. 37, no. 3, pp. 694–722, Jul. 2020.

[60] D. Audran, M. Andersen, M. Hansen, M. Andersen, T. Frederiksen, K. Hansen, D. Georgoulias, and E. Vasilomanolakis, ''Tick tock break the clock: Breaking CAPTCHAs on the darkweb,'' in *Proc. 19th Int. Conf. Secur. Cryptogr.*, 2022, pp. 357–365.

[61] A. Celestini, G. Me, and M. Mignone, ''Tor marketplaces exploratory data analysis: The drugs case,'' in *Global Security, Safety and Sustainability: The Security Challenges of the Connected World*, vol. 630, H. Jahankhani, A. Carlile, D. Emm, A. Hosseinian-Far, G. Brown, G. Sexton, and A. Jamal, Eds. Cham, Switzerland: Springer, 2016, pp. 218–229, doi: 10.1007/978-3-319-51064-4_18.

[62] A. Cuevas, F. Miedema, K. Soska, N. Christin, and R. van Wegberg, ''Measurement by proxy: On the accuracy of online marketplace measurements,'' in *Proc. 31st USENIX Secur. Symp. (USENIX Security)*, 2022, pp. 2153–2170. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/cuevas

[63] H. Zhang and F. Zou, ''A survey of the dark web and dark market research,'' in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1694–1705.

[64] Y. Fang, Y. Guo, C. Huang, and L. Liu, ''Analyzing and identifying data breaches in underground forums,'' *IEEE Access*, vol. 7, pp. 48770–48777, 2019.

[65] A. Zenebe, M. Shumba, A. Carillo, and S. Cuenca, ''Cyber threat discovery from dark web,'' in *Proc. 28th Int. Conf. Softw. Eng. Data Eng.*, 2019, pp. 163–174.

[66] N. Arnold, M. Ebrahimi, N. Zhang, B. Lazarine, M. Patton, H. Chen, and S. Samtani, ''Dark-net ecosystem cyber-threat intelligence (CTI) tool,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 92–97.

[67] M. Almukaynizi, E. Marin, E. Nunes, P. Shakarian, G. I. Simari, D. Kapoor, and T. Siedlecki, ''DARKMENTION: A deployed system to predict enterprise-targeted external cyberattacks,'' in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 31–36.

[68] S. Samtani, H. Zhu, and H. Chen, "Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (D-GEF)," *ACM Trans. Privacy Secur.*, vol. 23, no. 4, pp. 1–33, Nov. 2020.

[69] D. Georgoulias, J. M. Pedersen, M. Falch, and E. Vasilomanolakis, "A qualitative mapping of darkweb marketplaces," in *Proc. APWG Symp. Electron. Crime Res. (eCrime)*, Dec. 2021, pp. 1–15.

[70] R. van Wegberg, S. Tajalizadehkhoob, K. Soska, U. Akyazi, C. H. Gañán, B. Klievink, N. Christin, and M. van Eeten, "Plug and prey? Measuring the commoditization of cybercrime via online anonymous markets," in *Proc. 27th USENIX Secur. Symp. (USENIX Security)*. Berkeley, CA, USA: USENIX Association, 2018, pp. 1009–1026.

[71] F. Brenner, F. Platzer, and M. Steinebach, "Discovery of single-vendor marketplace operators in the tor-network," in *Proc. 16th Int. Conf. Availability, Rel. Secur. (ARES)*, Aug. 2021, pp. 1–10.

[72] F. Platzer, M. Schäfer, and M. Steinebach, "Critical traffic analysis on the tor network," in *Proc. 15th Int. Conf. Availability, Rel. Secur. (ARES)*, New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 1–10.

[73] M. Steinebach, M. Schäfer, A. Karakuz, K. Brandl, and Y. Yannikos, "Detection and analysis of tor onion services," in *Proc. 14th Int. Conf. Availability, Rel. Secur. (ARES)*, Aug. 2019, pp. 1–10.

[74] J. Martin and N. Christin, "Ethics in cryptomarket research," *Int. J. Drug Policy*, vol. 35, pp. 84–91, Sep. 2016.

[75] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, and A. R. Beresford, "Ethical issues in research using datasets of illicit origin," in *Proc. Internet Meas. Conf.*, Nov. 2017, pp. 445–462.

[76] J. T. Harviainen, A. Haasio, T. Ruokolainen, L. Hassan, P. Siuda, and J. Hamari, "Information protection in dark web drug markets research," in *Proc. 54th Hawaii Int. Conf. Syst. Sci.*, 2021, pp. 1–9.

[77] F. Platzer and A. Lux, "A synopsis of critical aspects for darknet research," in *Proc. 17th Int. Conf. Availability, Rel. Secur.*, Aug. 2022, pp. 1–8.

[78] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 213–224.

[79] A. Cantelo. (2016). *Deeplight: Shining a Light on the Dark Web*. [Online]. Available: https://www.slideshare.net/GavinOToole/deeplight-intelliagg

[80] V. Labrador and S. Pastrana, "Examining the trends and operations of modern dark-web marketplaces," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS PW)*, Jun. 2022, pp. 163–172.

[81] J. Broséus, D. Rhumorbarbe, M. Morelato, L. Staehli, and Q. Rossy, "A geographical analysis of trafficking on a popular darknet market," *Forensic Sci. Int.*, vol. 277, pp. 88–102, Aug. 2017.

[82] S. Samtani, M. Abate, V. Benjamin, and W. Li, "The dark web as a platform for crime: An exploration of illicit drug, firearm, CSAM, and cybercrime markets," in *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, T. J. Holt and A. M. Bossler, Eds. Cham, Switzerland: Springer, 2020, pp. 91–116, doi: 10.1007/978-3-319-78440-3.

[83] S. Samtani, M. Abate, V. Benjamin, and W. Li, "Cybersecurity as an industry: A cyber threat intelligence perspective," in *The Palgrave Handbook of International Cybercrime and Cyberdeviance*, T. J. Holt and A. M. Bossler, Eds. Cham, Switzerland: Springer, 2020, pp. 135–154, doi: 10.1007/978-3-319-78440-3.

[84] N. Ferry, T. Hackenheimer, F. Herrmann, and A. Tourette, "Methodology of dark web monitoring," in *Proc. 11th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, Jun. 2019, pp. 1–7.

[85] M. Bernaschi, A. Celestini, M. Cianfriglia, S. Guarino, F. Lombardi, and E. Mastrostefano, "Onion under microscope: An in-depth analysis of the tor web," *World Wide Web*, vol. 25, no. 3, pp. 1287–1313, May 2022.

[86] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "DarkBERT: A language model for the dark side of the internet," 2023, *arXiv:2305.08596*.

[87] H. Lawrence, A. Hughes, R. Tonic, and C. Zou, "D-miner: A framework for mining, searching, visualizing, and alerting on darknet events," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Oct. 2017, pp. 1–9.

[88] M. Graczyk and K. Kinningham. (2015). *Automatic Product Categorization for Anonymous Marketplaces*. [Online]. Available: https://cs229.stanford.edu/proj2015/184_report.pdf

[89] B. Akhgar, M. Gercke, S. Vrochidis, and H. Gibson, Eds., *Dark Web Investigation* (Security Informatics and Law Enforcement). Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-55343-2.

[90] N. Zhang, M. Ebrahimi, W. Li, and H. Chen, "A generative adversarial learning framework for breaking text-based CAPTCHA in the dark web," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2020, pp. 1–6.

[91] Y. Zhang, H. Gao, G. Pei, S. Luo, G. Chang, and N. Cheng, "A survey of research on CAPTCHA designing and breaking techniques," in *Proc. 18th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun., 13th IEEE Int. Conf. Big Data Sci. Eng. (TrustCom/BigDataSE)*, Aug. 2019, pp. 75–84.

[92] K. Csuka and D. Gaastra, "Breaking CAPTCHAs on the dark web," Dept. Syst. Netw. Eng., Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep., 2018. [Online]. Available: https://www.os3.nl/media/2017-2018/courses/rp1/p62report.pdf

[93] H. Weng, B. Zhao, S. Ji, J. Chen, T. Wang, Q. He, and R. Beyah, "Towards understanding the security of modern image captchas and underground captcha-solving services," *Big Data Mining Anal.*, vol. 2, no. 2, pp. 118–144, Jun. 2019.

[94] N. Zhang, M. Ebrahimi, W. Li, and H. Chen, "Counteracting dark web text-based CAPTCHA with generative adversarial learning for proactive cyber threat intelligence," *ACM Trans. Manage. Inf. Syst.*, vol. 13, no. 2, pp. 1–21, Jun. 2022.

[95] S. Nazah, S. Huda, J. H. Abawajy, and M. M. Hassan, "An unsupervised model for identifying and characterizing dark web forums," *IEEE Access*, vol. 9, pp. 112871–112892, 2021.

[96] B. L. Bullough, A. K. Yanchenko, C. L. Smith, and J. R. Zipkin, "Predicting exploitation of disclosed software vulnerabilities using open-source data," in *Proc. 3rd ACM Int. Workshop Secur. Privacy Anal.*, Mar. 2017, pp. 45–53.

[97] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Proc. 35th Annu. Symp. Found. Comput. Sci.*, 1994, pp. 124–134.

PHILIPP KÜHN is currently pursuing the Ph.D. degree with the Department of Computer Science, Technical University of Darmstadt, Chair of Science and Technology for Peace and Security (PEASEC). His primary research interests include information extraction from public data sources, emphasizing IT security, preparation, and processing, using natural language processing and deep learning methods. Furthermore, he conducts research on intergovernmental cooperation in for IT security.

KYRA WITTORF received the B.Sc. degree in computer science and the M.Sc. degree in IT security from the Technical University of Darmstadt, focusing on software security, distributed system security, and machine learning. Her master's thesis was on information gathering in the dark web.

CHRISTIAN REUTER received the Ph.D. degree. He is currently a Full Professor and the Dean of the Computer Science Department, Technical University of Darmstadt, leading the PEASEC Chair, which integrates computer science with peace and security research. He has published over 350 scientific works, received numerous awards, and leads several research projects. He is actively involved in university administration, major research initiatives, scientific societies, and conferences. His interdisciplinary research interests include cybersecurity and peace and conflict studies.